ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΧΗΜΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
Τομέας ΙΙ: Ανάλυσης, Σχεδιασμού και Ανάπτυξης Διεργασιών και Συστημάτων
Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής

Διπλωματική Εργασία

Ανάπτυξη διαδικτυακής βιβλιοθήκης μοντέλων μηχανικής μάθησης για την πρόβλεψη ιδιοτήτων νανοϋλικών

Παναγιώτα Ε. Κοτταρά

Επιβλέπων: Χαράλαμπος Σαρίμβεης, Καθηγητής Ε.Μ.Π.

Αθήνα 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΧΗΜΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
Τομέας ΙΙ: Ανάλυσης, Σχεδιασμού και Ανάπτυξης Διεργασιών και Συστημάτων
Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής

Διπλωματική Εργασία

Ανάπτυξη διαδικτυακής βιβλιοθήκης μοντέλων μηχανικής μάθησης για την πρόβλεψη ιδιοτήτων νανοϋλικών

Παναγιώτα Ε. Κοτταρά

Επιβλέπων: Χαράλαμπος Σαρίμβεης, Καθηγητής Ε.Μ.Π.

Αθήνα 2019

ΠΕΡΙΛΗΨΗ

Σκοπός της παρούσας εργασίας ήταν η δημιουργία εμιας διδικτυακής βιβλιοθήκης μοντέλων μηχανικής μάθησης για την πρόβλεψη ιδιοτήτων νανοϋλικών στις εφαρμογές ανοιχτού κώδικα Jaqpot Quattro και Jaqpot v5.

Το πρώτο κομμάτι ξεκινά με μια αναφορά στην εξέλιξη, την χρήση αλλα και τα πλεονεκτήματα των νανοϋλικών, συνεχίζει με μια εκτενή αναφορά στην μηχανική μάθηση και τους κυριότερους αλγορίθμους που χρησιμοποιεί και ολοκληρώνεται με τα μοντέλα QSAR όπου πραγματοποιείται μια σύντομη ιστορική αναδρομή και παρουσίαση της διαδιακασίας μοντελοποίησης και των βασικών συστατικών της.

Στο δεύτερο μέρος πραγματοποιείται αναλυτική παρουσίαση των δύο εφαρμογών μοντελοποίησης με αναφορά στις λειτουργίες τους αλλά και στο πως δημιουργούμε σύνολα δεδομένων και μοντέλα ενώ ολοκληρώνεται με την παρουσίαση του αποθετηρίου που δημιουργήθηκε.

Λέξεις κλειδιά: Νανοϋλικά, μοντελοποίηση, μηχανική μάθηση, αλγόριθμοι, QSAR, Jaqpot Quattro, Jaqpot v5

ABSTRACT

The purpose of this thesis was the creation of an online repository of machine learning models for predicting properties of nanomaterials in open source web platforms Jaqpot Quattro and Jaqpot v5.

The first part of the thesis starts with a reference to the evolution, use and benefits of nanomaterials, continues with a comprehensive reference to mechanical learning and the basic algorithms it uses and finishes with QSAR modeling including a brief historical review and presentation of the modeling process and its basic components.

In the second part of the thesis, there is a detailed presentation of the two Jaqpot GUIs, with reference to their functions as well as a presentation of the process of creating/ uploading datasets and models and in the last part there is a presentation of the repository created.

Keywords: Nanomaterials, Models, Machine Learning, Algorithms, QSAR, Jaqpot Quattro, Jaqpot v5

ΕΥΧΑΡΙΣΤΙΕΣ

Φτάνοντας στο τέλος αυτής της εργασίας με την οποία ολοκληρώνετα ένας έναν μακρύς ομολογουμένως κύκλος σπουδών στο ΕΜΠ, νιώθω την ανάγκη να ευχαριστήσω τους ανθρώπους που συνέβαλαν στην εκπονησή της.

Πρώτα απ'όλα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Χαράλαμπο Σαρίμβεη για την ανάθεση ενός τόσο ενδιαφέροντος θέματος και την υποστήριξή του καθ'όλη την διαρκεια εκπόνσης της εργασίας. Ακολούθως θα ήθελα να ευχαριστήσω τα μέλη που Εργαστηρίου Αυτόματης Ρύθμισης και Πληροφορικής με τα οποία συνεργάστηκα όλο αυτό το διάστημα. Τον Γιώργο Δρακάκη για τις πολύτιμες συμβουλές του στην αρχή της εργασίας, τον Παντελή Καρατζά για την υποστήριξη στα τεχνικά ζητήματα που προέκυπταν και τον Φίλιππο Δογάνη για την βοήθεια και την υποστήριξή του το τελευταίο και πιο κρίσιμο διάστημα.

Ακολούθως θα ήθελα να ευχαριστήσω την κυρία Μαργαρίτα Μπεάζη – Κατσιώτη, Καθηγήτρια της Σχολής Χημικών Μηχανικών Ε.Μ.Π και τον κύριο Φώτη Τσόπελα, Λέκτορα της Σχολής Χημικών Μηχανικών Ε.Μ.Π, μέλη της εξεταστικής επιτροπής για τον χρόνο που αφιέρωσαν στην ανάγνωση της παρούσας εργασίας.

Στην συνέχεια θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στην Κωνσταντίνα, που εκτός από εξαιρετική συνεργάτης είναι και εξαίρετη φίλη, για την υποστήριξη της το τελευταίο διάστημα.

Τέλος το μεγαλύτερο ευχαριστώ το οφείλω στους ανθρώπους που αποτελούν φάρους στην πορεία της ζωής μου. Την αδερφή μου Ρούλα, που είναι πάντα δίπλα μου και δεν σταμάτησε ποτέ να πιστεύει σε μένα και τους γονείς μου Βαγγέλη και Δήμητρα για όλες τις θυσίες που έκαναν για να φτάσουμε εγώ και η αδερφή μου ως εδώ. Εύχομαι μια μέρα να καταφέρω να τους κάνω περήφανους.

Παναγιώτα Κοτταρά

Ιούνιος 2019

Στους γονείς μου, Βαγγέλη και Δήμητρα

Στην αδερφή μου, Ρούλα

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Εισαγωγή

Τα τελευταία χρόνια το βλέμμα της επιστημονικής και όχι μόνο κοινότητας έχει στραφεί προς έναν νέο τεχνολογικό τομέα, αυτόν την νανοτεχνολογίας, που αποτελεί την μετεξέλιξη της μικροτεχνολογίας και βρίσκει εφαρμογές σε ολοένα και περισσότερα πεδία. Η ανάπτυξη της νανοτεχνολογίας προσφέρει τη δυνατότητα αξιοποίησης υφιστάμενων, φυσικών, νανοϋλικών αλλά και σύνθεσης και χρήσης νέων, κατασκευασμένων νανοϋλικών (ENM) και συμβάλλει καθοριστικά στην επίλυση σημαντικών θεμάτων σε διάφορους τομείς της επιστήμης. Τα κατασκευασμένα νανοϋλικά, λόγω των ιδιοτήτων που παρουσιάζουν και του εξαιρετικά μικρού τους μέγεθος έχουν ποικίλλες εφαρμογές με παραδείγματα να αποτελούν η ιατρική, οι τομείς της ενέργειας (παραγωγή, αποθήκευση, εξοικονόμηση), η προστασία του περιβάλλοντος, η αυτοκινητοβιομηχανία, η ηλεκτρονική, η βιομηχανία καλλυντικών, τροφίμων και ποτών. Κρίνοντας λοιπόν από το ευρύ φάσμα εφαρμογών τους, γίνεται εύκολα αντιληπτή η επιτακτική ανάγκη ορθής εφαρμογής τους με σκοπό την αξιοποίηση των πλεονεκτημάτων τους αλλά και την αποφυγή των κινδύνων που εγκυμονεί η χρήση τους.

Το εξαιρετικά μικρό μέγεθος των νανοϋλικών σε συνδυασμό με τις ιδιότητές τους που είναι απόρροια των φυσικοχημικών χαρακτηριστικών τους, εκτός από τις αναρίθμητες δυνατότητες αξιοποίησής τους, έχει δημιουργήσει και έντονη ανησυχία για τους κινδύνους που μπορεί να ελλοχεύουν. Για την μελέτη των ανεπιθύμητων ιδιοτήτων και των αρνητικών επιπτώσεων της χρήσης των νανοϋλικών στους ζωντανούς οργανισμούς και στο περιβάλλον, πραγματοποιείται έντονη πειραματική δραστηριότητα. Το πλήθος όμως

των νανοϋλικών και των παραλλαγών τους που προκύπτουν από διαφορετικά μεγέθη, σχήματα, επικαλλύψεις κλπ., καθιστούν απαγορευτική την πειραματική τους μελέτη, λόγω περιορισμών χρόνου αλλά και κόστους. Η ανάπτυξη της επιστήμης των υπολογιστών προσφέρει μια ενναλλακτική προσέγγιση που βασίζεται στις υπολογιστικές (in silico) μελέτες, μελέτες δηλαδή που πραγματοποιούνται με την χρήση μαθηματικών μοντέλων προσομοίωσης. Οι in silico τεχνικές, εκτός από το συγκρίσιμα χαμηλότερο κόστος και χρόνο, συνεισφέρουν στη μείωση της χρήσης πειραματόζωων για ερευνητικούς σκοπούς, κάτι που άλλωστε είναι ιδιαίτερα επιθυμητό βάσει της Ευρωπαϊκής οδηγίας REACH.

Μια κατηγορία των in silico μεθόδων που χρησιμοποιείται ευρύτατα τελευταία, είναι τα μοντέλα που συσχετίζουν τη δομή με την δραστικότητα (Structure Activity Relationships, SAR) και τα μοντέλα ποσοτικής σχέσης δομής – δραστικότητας (Quantitative Structure Activity Relationships, QSAR). Πρόκειται για μαθηματικά μοντέλα ικανά να προβλέψουν φυσικοχημικές και βιολογικές ιδιότητες ενώσεων, με βάση τη δομή τους και περιγράφονται από την μαθηματική συναρτησιακή σχέση: **Βιολογική Δράση = f (Δομή).** Οι ποσοτικές σχέσεις δομής – δραστικότητας βασίζονται στην υπόθεση ότι η δομή ενός μορίου σχετίζεται με τα χαρακτηριστικά εκείνα που είναι υπεύθυνα για τις φυσικές, χημικές ή βιολογικές ιδιότητες. Με τη χρήση των σχέσεων-μοντέλων QSAR καθίσταται δυνατός ο προσδιορισμός της βιολογικής συμπεριφοράς, ιδιότητας ή δραστικότητας μιας νέας ουσίας με βάση τη μοριακή δομή άλλων παρόμοιων ουσιών, των οποίων η αντίστοιχη ιδιότητα έχει ήδη εκτιμηθεί.

Η αυξημένη αποδοχή που χαίρουν τα μοντέλα QSAR απο ολοένα και μεγαλύτερο ποσοστό της επιστημονικής κοινότητας δημιουργεί την ανάγκη δημιουργίας υποδομών κατάλληλων να διαχειριστούν και να φιλοξενούν δεδομένα τα οποία θα χρησιμοποιηθούν για την ανάπτυξη προγνωστικών μοντέλων. Αυτή την ανάγκη φιλοδοξούν να καλύψουν οι δύο πλατφόρμες Jaqpot Quattro και η μετεξέλιξή της Jaqpot v5 που αναπτύσσονται από τη Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου. Πρόκειται για δύο ολοκληρωμένες υπολογιστικές εφαρμογές ανοιχτού κώδικα ειδικά σχεδιασμένες για να παρέχουν στον χρήστη πληθώρα

δυνατοτήτων όπως π.χ. η εισαγωγή, επιλογή, προετοιμασία και επεξεργασία δεδομένων ώστε να χρησιμοποιηθούν στην μοντελοποίηση. Οι δύο πλατφόρμες ενσωματώνουν αλγορίθμους στατιστικής, εξόρυξης δεδομένων και μηχανικής μάθησης με ταυτόχρονη δυνατότητα του χρήστη να δημιουργήσει τις δικές του ροές εργασίας για την κατασκευή μοντέλων σχέσεων νανο-ποσοτικής δομής-δραστηριότητας (nanoQSAR modeling) τα οποία μπορεί στην συνέχεια να αξιολογήσει μέσω των παρεχόμενων δυνατοτήτων επικύρωσης (split-, cross- and external validation). Όλα αυτά πραγματοποιούνται σε ένα περιβάλλον φιλικό προς τον χρήστη που καθιστά τη χρήση των εργαλείων ακόμα και από μη εξοικειωμένους με τους Η/Υ χρήστες ιδιαίτερα απλή διαδικασία, ενώ η αρχιτεκτονική δομή των εργαλείων επιτρέπει την διαρκή ανάπτυξη και βελτίωσή τους τόσο από τους προγραμματιστές της Μονάδας όσο και από ολόκληρη την κοινότητα των χρηστών. Οι δύο πλατφόρμες προσφέρονται στην επιστημονική κοινότητα ως πλατφόρμες ανοιχτού κώδικα, οι οποίες δεν περιορίζονται στην ανάπτυξη και την ανταλλαγή προγνωστικών μοντέλων, αλλά περιλαμβάνουν επίσης λειτουργίες που υποστηρίζουν και αξιολογούν τη συνεργασία μεταξύ των μοντέλων και των υπολογιστικών ερευνητικών ομάδων.

Δεδομένου ότι τα μοντέλα QSAR για την πρόβλεψη ιδιοτήτων των νανοϋλικών χρησιμοποιούνται από περισσότερους επιστήμονες και ερευνητές διαφορετικού γνωστικού υπόβαθρου που ενδεχομένως να μην έχουν την γνώση να δημιουργήσουν ένα μοντέλο, σε συνδυασμό με το ότι οι δύο πλατφόρμες παρέχουν μια ολοκληρωμένη υποδομή με δυνατότητες αποθήκευσης, ανταλλαγής και αναζήτησης δεδομένων, γενήθηκε η ιδέα της δημιουργίας μιας διαδικτυακής βιβλιοθήκης μοντέλων μηχανικής μάθησης για την πρόβλεψη ιδιοτήτων των νανοϋλικών τα οποια θα είναι προσβάσιμα από την επιστημονική κοινότητα για χρήση στην έρευνα. Στο πλαίσιο της εργασίας αυτής δημιουργήθηκε ένα ηλεκτρονικό αποθετήριο υλοποιημένων μοντέλων της βιβλιογραφίας το οποίο είναι διαθέσιμο και στις δύο εκδοχές του Jaqpot και είναι προσβάσιμο στις ηλεκτρονικές διευθύνσεις https://jaqpot.org/ (Jaqpot Quattro) και https://app.jaqpot.org/ (Jaqpot v5). Τα μοντέλα είναι διαθέσιμα ως έτοιμες και εύχρηστες εφαρμογές ιστού που συνοδεύονται από μετα-πληροφορίες που επιτρέπουν στους χρήστες να αναζητήσουν και να βρουν το πιο κατάλληλο μοντέλο για τις συγκεκριμένες ανάγκες τους.Τα μοντέλα που υλοποιήθηκαν

επιλέχθηκαν από την βιβλιογραφία με βασικότερο κριτήριο τη διαθεσιμότητα των απαραίτητων δεδομένων. Στόχος είναι η συνεχής ενημέρωση της βιβλιοθήκης με νέα μοντέλα που θα είναι διαθέσιμα μέσω ενός ή και των δύο GUI της υπολογιστικής πλατφόρμας.

1. Νανοϋλικά & Νανοτεχνολογία

1.1. Γενικά

Τα τελευταία χρόνια όλο και συχνότερα ακούμε γύρω μας τις λέξεις νανοτεχνολογία, νανοϋλικά και νανοσωματίδια. Είναι γεγονός πως ο τομέας της νανοτεχνολογίας τελευταία καταλαμβάνει ολοένα και μεγαλύτερο τμήμα από την «πίτα» της επιστήμης και της έρευνας. Παρά το γεγονός πως η άνθηση στον τομέα της νανοτεχνολογίας είναι πρόσφατη, οι πράγματι σπουδαίες ιδιότητες των νανοσωματιδίων ήταν γνωστές αρκετά χρόνια πίσω. Το 1959 ο καθηγητής φυσικής Richard Feynman, έδωσε τη διάλεξή με θέμα: «There's plenty of room at the bottom» η οποία ήρθε να ταράξει τα νερά της επιστημονικής κοινότητας. Τότε έθεσε το ερώτημα «*Why cannot we write the entire 24 volumes of the Encyclopedia Brittanica on the head of a pin?*» και μίλησε για την πιθανότητα διάταξης των ατόμων βάση του τρόπου που ο επιστήμονας επιθυμεί, εννοώντας πως ο Χημικός θα μπορούσε να δημιουργήσει οποιαδήποτε χημική ουσία διατάσσοντας τα άτομα με τρόπο που θα του υποδείκνυε ο Φυσικός. [10]

Η λέξη νανοσωματίδιο (στα αγγλικά nanoparticle) αποτελεί σύνθεση του προθέματος «νάνο-» που συνδέεται ετυμολογικά με την ελληνική λέξη νάνος και προσδιορίζει κάτι πολύ μικρό και την λατινική λέξη «particulum» που σημαίνει σωματίδιο. Το νάνο ως πρόθεμα μονάδας μέτρησης δηλώνει τάξη μεγέθους 10^{-9} και μπορεί να αφορά μονάδα μέτρησης του μήκους, του όγκου, της μάζας, του βάρους ή και του χρόνου. Στην νανοτεχνολογία χρησιμοποιείται σχεδόν αποκλειστικά ο όρος νανόμετρο, μια υποδιαίρεση του μέτρου. Για την ακρίβεια το νανόμετρο ισούται με ένα (1) δισεκατομμυριοστό του μέτρου (1 nm=10^{-9} m) ή αλλιώς ένα εκατομμυριοστό του χιλιοστόμετρου (1nm=10^{-6} mm). Έτσι όταν αναφερόμαστε σε νανοϋλικά και νανοσωματίδια αναφερόμαστε στα υλικά, ή τα σωματίδια αντίστοιχα, που έχουν τουλάχιστον μία διάσταση μεταξύ 1 και 1000 νανομέτρων (1-1000 nm) αν και ο συνηθέστερος ορισμός της νανοκλίμακας είναι μεταξύ 1 και 100 νανομέτρων (1-100 nm). [4] Θα μπορούσαμε λοιπόν να πούμε κατά μια έννοια ότι η νανοτεχνολογία αποτελεί μια αναμενόμενη εξέλιξη αν αναλογιστούμε τη στροφή που έχει πραγματοποιηθεί προς την «σμίκρυνση» τις τελευταίες δεκαετίες, με την εξέλιξη της επιστήμης των υπολογιστών να διαδραματίζει κυρίαρχο ρόλο σε αυτό. Δεν είναι άλλωστε

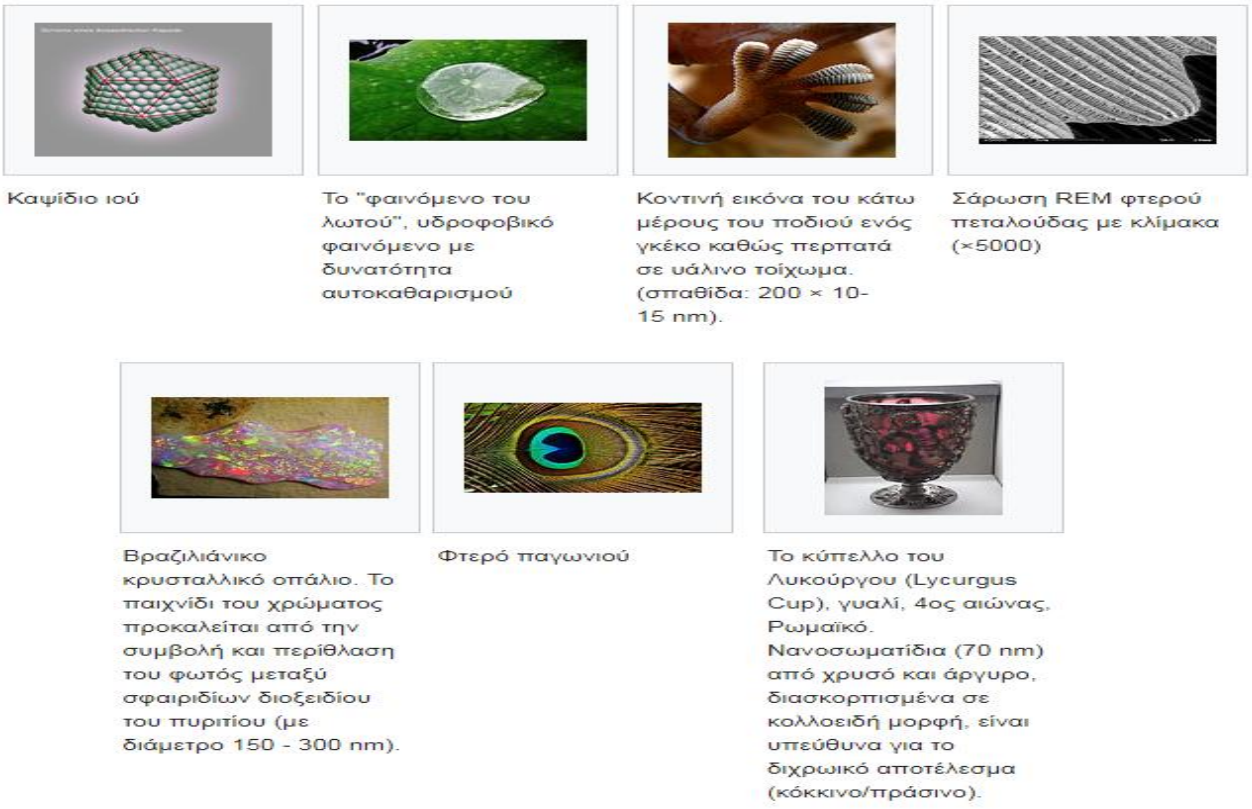
τυχαίο πως προϊόντα με διαστάσεις στην κλίμακα του μικρομέτρου (μm), δηλαδή ένα εκατομμυριοστό του μέτρου (10^{-6}) χρησιμοποιούνται ευρέως εδώ και χρόνια σε διάφορους τομείς της βιομηχανίας (π.χ. αυτοκινητοβιομηχανία και αεροναυπηγική). [21]

Τα νανοϋλικά χωρίζονται, μεταξύ άλλων κατηγοριοποιήσεων, σε φυσικά και συνθετικά.

1.2. Φυσικά Νανοϋλικά

Ως φυσικά νανοϋλικά ορίζονται αυτά που απαντώνται στη φύση ή δημιουργούνται από φυσικές διεργασίες όπως οι εκρήξεις ηφαιστείων ή οι πυρκαγιές. Συχνά τα βιολογικά συστήματα χαρακτηρίζονται από φυσικά λειτουργικά νανοϋλικά (φυσικά βιολογικά νανοϋλικά). Παραδείγματα αποτελούν η δομή των καψιδίων ιών, οι κρύσταλλοι κεριού που καλύπτουν έναν λωτό ή ένα φύλλο τροπαίολου, αράχνης και μεταξιού του τετράνυχου (spider-mite silk) [22], φυσικά κολλοειδή (γάλα, αίμα), υλικά όπως (δέρμα, νύχια, ράμφη, φτερά, κέρατα, τρίχα), χαρτί, βαμβάκι, μάργαρο, κοράλλια κι ακόμα οι δικοί μας οστεώνες είναι όλοι φυσικά οργανικά νανοϋλικά ().

Στον αντίποδα φυσικά ανόργανα νανοϋλικά υπάρχουν μέσα από την κρυσταλλική ανάπτυξη σε διάφορες χημικές συνθήκες του φλοιού της γης. Παραδείγματος χάρη, οι άργιλοι εμφανίζουν σύμπλοκες νανοδομές λόγω της ανισοτροπίας της υποκείμενης κρυσταλλικής τους δομής και η ηφαιστειακή δραστηριότητα μπορεί να προκαλέσει τα οπάλια, που είναι ένα στιγμιότυπο μιας φυσικής εμφάνισης φωτονικών κρυστάλλων λόγω της δομής τους σε νανοκλίμακα. Οι πυρκαγιές αντιπροσωπεύουν ιδιαίτερα σύνθετες αντιδράσεις και μπορούν να παράξουν χρωστικές, τσιμέντο, ατμίζον διοξείδιο του πυριτίου (fumed silica) κλπ. [54]



Εικόνα 1: Παραδείγματα φυσικών νανοϋλικών
Πηγή: https://el.wikipedia.org

1.3. Συνθετικά νανοϋλικά

Με τον όρο συνθετικά νανοϋλικά περιγράφονται τα νανοσωματίδια εκείνα που συντίθενται στο εργαστήριο και χρησιμοποιούνται ευρύτατα από την παγκόσμια βιομηχανία σε προϊόντα και διαδικασίες λόγω των ιδιαίτερα σημαντικών ιδιοτήτων τους.[29] με μια από τις σημαντικότερες να είναι η αναλογία επιφάνειας προς όγκο (surface-to-volume ratio) και η οποία αυξάνεται καθώς μειώνεται η διάμετρος του νανοσωματιδίου. Αυτό πρακτικά σημαίνει πως η αναλογία των ατόμων που συνθέτουν το νανοσωματίδιο και που βρίσκονται στην επιφάνεια του αυξάνεται, με αποτέλεσμα την αύξηση των ακόρεστων δεσμών ή των ελεύθερων θέσεων σύνταξης που έχει ως συνέπεια την αύξηση της δραστικότητας του σωματιδίου [43] δίνοντας έτσι αναρίθμητες δυνατότητες, με πιο κοινή τη δημιουργία νέας γενιάς καταλυτών.

Τα συνθετικά νανοϋλικά μπορεί να είναι αποτέλεσμα σύνθεσης μετάλλων, υλικών ημιαγωγών, οξειδίων μετάλλων, ή οργανικών υλικών. Στην Εικ. 2 φαίνεται μια κατηγοριοποίηση σύνθετων νανοσωματιδίων με βάση τα φυσικά και χημικά χαρακτηριστικά τους.

Carbonic	Metal oxides	Semiconductors	Metals
fullerenes nanotubes Carbon Black	silicon dioxide (SiO_2) titanium dioxide (TiO_2) aluminum oxide (Al_2O_3) iron oxide (Fe_2O_3) or (Fe_3O_4), zinc oxide (ZnO)	cadmium-tellurite (CdTe) silicon (Si) indium phosphide (InP or InGaP)	gold (Au) silver (Ag) iron (Fe) cobalt (Co)

Εικόνα 2: Ομαδοποίηση συνθετικών νανοσωματιδίων
Πηγή: Raab, Simkó, Gazsó, & Fiedeler, 2011

Η ανακάλυψη νέων δομών του άνθρακα μπορεί να θεωρηθεί σταθμός για την επιστήμη της νανοτεχνολογίας. Το αλλόμορφο του άνθρακα γνωστό ως φουλερένιο (fullerene, C_{60}) που ανακαλύφθηκε το 1985 από τους Curl Krot και Smalley και πήρε το όνομά του από τον αρχιτέκτονα Buckminster Fullerene άνοιξε νέους δρόμους στην νανοτεχνολογία. Το φουλερένιο έχει σφαιρική δομή και αποτελείται από 60 ή παραπάνω άτομα άνθρακα, ενώ η διάμετρός του κυμαίνεται μεταξύ 0,7nm – 1,5nm. Το 1991 το φουλερένιο C60 βραβεύτηκε ως σωματίδιο της χρονιάς και από τότε υπήρχαν ενδείξεις για δυνητική χρήση στην ιατρική λόγω των φυσικών και χημικών ιδιοτήτων τους. Σήμερα η χρήση του C60 αποτελεί καινοτόμο μέθοδο για την θεραπεία βαρύτατων ασθενειών όπως ο καρκίνος αλλά και οι νευροεκφυλιστικές διαταραχές.

Στα πλαίσια της συγκεκριμένης εργασίας όταν αναφερόμαστε σε συνθετικά νανοϋλικά θα χρησιμοποιούμε τον όρο «κατασκευασμένα νανοϋλικά» (Engineered nanomaterials, ENM) που σύμφωνα με το ISO/TS 80004-1:2010 περιγράφει υλικά σε νανο-κλίμακα που έχουν κατασκευαστεί από τον άνθρωπο κα είναι"Σχεδιασμένα για συγκεκριμένο σκοπό ή λειτουργία"

1.4. Εφαρμογές των νανοϋλικών

Ο τομέας της νανοτεχνολογίας λόγω των αναρίθμητων δυνατοτήτων που προσφέρει έχει στρέψει επάνω της τα βλέμματα όλης της επιστημονικής κοινότητας. Το γεγονός πως μέσω της ανάπτυξή της έδωσε την δυνατότητα αξιοποίησης υφιστάμενων, φυσικών, νανοϋλικών αλλά και σύνθεσης και χρήσης νέων, συνθετικών, αποτέλεσε συνέβαλε καθοριστικά στην επίλυση σημαντικών και δυσεπίλυτων για χρόνια θεμάτων που ταλάνιζαν διάφορους τομείς της επιστήμης. Εξαιτίας αυτής της συμβολής της η νανοτεχνολογία βρίσκεται πρώτη στην λίστα ανάμεσα στους τομείς της επιστήμης που προσελκύουν δημόσια χρηματοδότηση για την έρευνα και την εξέλιξή της. Πρόκειται επίσης για έναν πολυεπιστημονικό τομέα αφού συνδυάζει τις επιστήμες της ιατρικής, της βιολογίας, της χημείας, της μηχανικής, των υλικών και τα τελευταία χρόνια των μαθηματικών και τις πληροφορικής όσον αφορά την μοντελοποίηση των νανοϋλικών, που αποτελεί και κομμάτι της συγκεκριμένης εργασίας. Γίνεται λοιπόν εύκολα αντιληπτό ότι η συνεργασία επιστημόνων διαφόρων ειδικοτήτων και διαφορετικού επιστημονικού υπόβαθρου οδηγεί σε νέες διαπιστώσεις και ανακαλύψεις που έρχονται να αντικαταστήσουν τις παλιές, «απαρχαιωμένες» προσεγγίσεις σε μια εποχή που η εξέλιξη στις θετικές και τεχνολογικές επιστήμες είναι ραγδαία.

1.4.1. Εφαρμογές των νανοϋλικών στην Ιατρική

Ένας από τους πρώτους τομείς στους οποίους βρήκε εφαρμογή η γνώση περί νανοϋλικών και νανοτεχνολογίας είναι η ιατρική. Τα νανοϋλικά και τα νανοσωματίδια χρησιμοποιούνται σε πολλές εκφάνσεις της ιατρικής επιστήμης εδώ και αρκετά χρόνια με το «breakthrough» ωστόσο να είναι αρκετά πρόσφατο, όσο πρόσφατος είναι και ο τομέας της νανοϊατρικής. Ως νανοϊατρική ορίζεται η εφαρμογή της νανοτεχνολογίας στην επιστήμη της ιατρικής, κάτι που καθίσταται δυνατό με την αξιοποίηση των φυσικών, χημικών και βιολογικών ιδιοτήτων των νανοδομημένων υλικών και στο γεγονός πως το σημαντικά μικρότερο μέγεθός τους εν συγκρίσει με το μέγεθος των κυττάρων και των

βιολογικών μορίων τα καθιστά πιο εύκολα διαχειρίσιμα και διευκολύνει την αλληλεπίδρασή τους με τους έμβιους οργανισμούς.[38]

Η χρήση των νανοϋλικών από την ιατρική στοχεύει στην βελτίωση των μεθόδων που χρησιμοποιούνται τόσο στην διάγνωση όσο και στη θεραπεία. Η ανάπτυξη της νανοτεχνολογίας έχει δώσει μεγάλη ώθηση στην ανάπτυξη της **διαγνωστικής ιατρικής** προσφέροντας αναρίθμητες νέες δυνατότητες διάγνωσης είτε πρόκειται για in vitro, ex vivo ή in vivo διάγνωση. Χαρακτηριστικά αξίζει να αναφέρουμε ότι έχουν ήδη αναπτυχθεί συσκευές διάγνωσης στην κλίμακα του νανομέτρου (chips) ικανές ακόμα και να αναγνωρίσουν τμήματα του ανθρώπινου DNA και των εκφρασμένων πρωτεινών τους (in vitro). Γενικά θα μπορούσαμε να πούμε ότι η διαγνωστική ιατρική μέσω της νανοτεχνολογίας στοχεύει στην ανάπτυξη νέων νανοσυσκευών με σκοπό:

- την ενσωμάτωση της διεργασίας προετοιμασίας του δείγματος στις συσκευές.
- την κατά το δυνατόν σμίκρυνση των χρησιμοποιούμενων συσκευών διάγνωσης που θα έχει ως αποτέλεσμα την απαίτηση μικρότερου όγκου των βιολογικών δειγμάτων.
- την ανάπτυξη διαγνωστικών συσκευών με σκοπό την πραγματοποίηση όσο το δυνατόν περισσότερων λειτουργιών σε μια συσκευή σε συνδυασμό με τη δυνατότητα συλλογής δεδομένων από απόσταση.
- την χρήση νανοσωματιδίων που προσφέρουν την δυνατότητα πρώιμης διάγνωσης παθολογιών καθιστώντας έτσι δυνατή την έγκαιρη αντιμετώπισή τους.

Εκτός όμως από τη διαγνωστική ιατρική, μεγάλη είναι η συμβολή των νανοϋλικών και στην **θεραπευτική ιατρική** όπου τα νανοϋλικά διαδραματίζουν πρωταγωνιστικό ρόλο με την χρήση τους στην ελεγχόμενη αποδέσμευση των φαρμάκων η οποία θα μπορούσε να χαρακτηριστεί απόρροια της ανάπτυξης συνθετικών νανοϋλικών που σκοπό έχουν την στοχευμένη απόδοση φαρμάκων και βιομορίων αφού είναι σε θέση να εντοπίζουν και να στοχεύουν συγκεκριμένα μόρια των παθολογικών κυττάρων. Αυτό γίνεται ευκολότερα κατανοητό αν αναλογιστεί κανείς πως η μη φυσιολογική λειτουργία των παθολογικών περιοχών οδηγεί σε ενισχυμένη διαπερατότητα και κατακράτηση, γεγονός που διευκολύνει τη συσσώρευση των νανοσωματιδίων σε εστίες μόλυνσης, φλεγμονές ή καρκινικούς όγκους. Η σύγχρονη ιατρική έχει στραφεί προς αυτή την κατεύθυνση στοχεύοντας στην

χρήση νανοϋλικών ικανών να προσροφούν φαρμακευτικές ενώσεις, εξασφαλίζοντας παράλληλα την προστασία τους από κάθε πιθανή χημική ή ενζυματική αποικοδόμηση και να τις αποδεσμεύουν στοχευμένα στις προς θεραπεία περιοχές. Αυτό μπορεί να επιτευχθεί αξιοποιώντας τα σημαντικά πλεονεκτήματα που παρουσιάζει η χρήση «νανομεταφορέων» φαρμάκων έναντι άλλων ειδών μεταφορέων. Σε αυτά συγκαταλέγεται το μικρό τους μέγεθος, η διαλυτότητά τους που ενισχύει την βιοδιαθέσιμότητά τους καθώς και η δυνατότητα στοχευμένης χορήγησης, μειωμένης τοσξικότητας θεραπείας, απευθείας σε συγκεκριμένους στόχους (πχ καρκινικούς όγκους) μέσω σημάτων τα οποία μπορούν να βασίζονται στην θερμοευαισθησία ή κάποια άλλη ιδιότητα.

1.4.2. **Εφαρμογές των νανοϋλικών στο περιβάλλον**

Η νανοτεχνολογία και η χρήση των νανοϋλικών όπως έχει ήδη αναφερθεί βρίσκει εφαρμογή όλο και σε περισσότερους τομείς της επιστήμης και της καθημερινής ζωής. Από τους τομείς αυτούς, δεν θα μπορούσε να λείπει και το περιβάλλον, και κυρίως ο τομέας της ενέργειας. Η ραγδαία αύξηση του πληθυσμού συνεπάγεται αύξηση της απαιτούμενης κατά κεφαλήν ενέργειας. Η αυξημένη αυτή κατανάλωση έχει δημιουργήσει σημαντικές μεταβολές στο κλίμα, την βιοποικιλότητα, την ποιότητα του νερού, του αέρα και της ζωής γενικότερα. Με την ανάπτυξη λοιπόν της νανοτεχνολογίας και των νανοϋλικών έχουν πραγματοποιηθεί σημαντικά άλματα στους τομείς κυρίως της παραγωγής αλλά και της αποθήκευσης και εξοικονόμησης ενέργειας τα οποία είναι σύμφωνα με την αειφόρο ανάπτυξη που προωθείται τον 21^{ο} αιώνα τον οποίο διανύουμε.

Ειδικότερα στην **παραγωγή** ενέργειας, ιδιαίτερα μετά την στροφή που έχει πραγματοποιηθεί προς την αξιοποίηση της ηλιακής ενέργειας, τα νανοϋλικά χρησιμοποιούνται τόσο στην ανάπτυξη συστημάτων φωτοβολταϊκών, τα οποία λόγω και του μικρότερου, συγκριτικά με τα προγενέστερα, μεγέθους τους είναι κατάλληλα ακόμα και για οικιακή χρήση όσο και στις κυψέλες καυσίμων.[40] Χαρακτηριστικό είναι το πρόσφατο παράδειγμα του ερευνητή Yang Yan από το Πανεπιστημίου Κεντρικής Florida[53]

ο οποίος σύμφωνα με δημοσίευσή[20] του το φθινόπωρο του 2017 επινόησε ένα νέο υβριδικό νανοϋλικό ικανό να αξιοποιεί την ηλιακή ενέργεια για την παραγωγή υδρογόνου από θαλασσινό νερό. Το παραχθέν υδρογόνο θα χρησιμοποιείται για την τροφοδότηση κυψελών καυσίμων και μολονότι η χρήση υδρογόνου για την τροφοδότηση κυψελών καυσίμου ήταν γνωστή, το υψηλό κόστος της ηλεκτρικής ενέργειας που απαιτείτο για την όλη διαδικασία καθιστά ακόμη σημαντικότερη την ανακάλυψη του Yan και της ομάδας του καθώς μελλοντικά θα μπορούσε να οδηγήσει σε μια νέα πηγή καθαρού καυσίμου, μειώνοντας τις απαιτήσεις σε ορυκτά καύσιμα και κατά συνέπεια την επιβάρυνση στο περιβάλλον. Επειδή όμως εκτός από την παραγωγή απαιτείται σύνεση και στην **αποθήκευση** της ενέργειας, σημαντική είναι η συμβολή των νανοϋλικών στην δημιουργία φιλικών προς το περιβάλλον μπαταριών και πυκνωτών ενώ η αξιοποίηση των νανοϋλικών στους τομείς των μονώσεων αλλά και του φωτισμού επιβεβαιώνουν πως μπορούν να συμβάλλουν και στην **εξοικονόμηση** ενέργειας.

Εκτός όμως από τον τομέα της ενέργειας τα οφέλη της νανοτεχνολογίας για το περιβάλλον εντοπίζονται και σε άλλους τομείς όπως την παραγωγή φιλικών προς το περιβάλλον φυτοφαρμάκων και λιπασμάτων, με συνέπεια τη μείωση της ρύπανσης του υδροφόρου ορίζοντα, τη χρήση νανοσωματιδίων για την βελτίωση της ποιότητας του ατμοσφαιρικού αέρα (ιδιαίτερα χρήσιμο σε βιομηχανικές περιοχές), ενώ τέλος θα ήταν παράλειψη να μην αναφέρουμε την δημιουργία νέων eco friendly υλικών, όπως είναι τα αντιδιαβρωτικά των μνημείων, ή οι αυτοκαθαριζόμενες επιστρώσεις των κτιρίων που έχουν ήδη αρχίσει να χρησιμοποιούνται, για παράδειγμα στο Μεξικό[21], για την αποτροπή των βανδαλισμών των κτιρίων (graffity) και έχουν αποδειχθεί ιδιαίτερα χρήσιμα μέσα για την αναβάθμιση των περιοχών ενώ παράλληλα μειώνουν το περιβαλλοντικό αποτύπωμα συγκριτικά με τα υλικά που χρησιμοποιούνταν παλιότερα και περιέχουν χρώμιο, κάδμιο και άλλες επικίνδυνες τοξικές ουσίες, τη χρήση των οποίων θέλει να περιορίσει η Ευρωπαϊκή Ένωση.

1.4.3. **Εφαρμογές των νανοϋλικών σε άλλους τομείς**

Μπορεί η ιατρική και το περιβάλλον να είναι οι μείζονες τομείς εφαρμογής των νανοϋλικών όμως είναι πολύ δύσκολο να εντοπιστεί τομέας που δεν έχει κάνει χρήση των δυνατοτήτων που παρέχει ο νέος αυτός αναπτυσσόμενος κλάδος της επιστήμης. Ξεκινώντας από την **ηλεκτρονική** όπου συνέβαλαν στην δημιουργία αισθητήρων πάσης φύσεως, την επιστήμη των ηλεκτρονικών υπολογιστών, τηλεπικοινωνίες, την οπτική με τη δημιουργία ανακλαστικών, αντι-ανακλαστικών επικαλύψεων, μέχρι τις εφαρμογές μεγάλης κλίμακας όπως οι εύκαμπτες ηλεκτρονικές διατάξεις και οι επίπεδες οθόνες απεικόνισης. Στην **αυτοκινητοβιομηχανία** η χρήση των νανοϋλικών εκτείνεται από την χρήση τους για την δημιουργία ελαφρύτερων αμαξωμάτων, μέχρι την παραγωγή καθαρότερων καυσίμων με χαμηλότερες τοξικές εκπομπές και βελτιωμένων λιπαντικών, τις ανθεκτικές στις γρατσουνιές επιστρώσεις, και την δημιουργία ελαφρύτερων και ισχυρότερων κινητήρων.

Η εργασία στη νανοκλίμακα δεν είναι νέα στις εταιρείες τροφίμων και ποτών. Πολλά τρόφιμα και τα ποτά περιέχουν φυσικά συστατικά που έχουν μέγεθος νανοκλίμακας ενώ φυσικά νανοσωματίδια εμπλέκονται εδώ και χρόνια και στην επεξεργασία τροφίμων. Τελευταία η ανάπτυξη της νανοτεχνολογίας έδωσε στις εταιρείες τροφίμων και ποτών τη δυνατότητα να παράγουν νέα αρώματα, γεύσεις προσδίδοντας άλλες φυσικές, οπτικές και αισθητικές διαστάσεις στα υπάρχοντα προϊόντα αλλά και νέα συμπληρώματα διατροφής όπως π.χ. βιταμίνες που προσφέρουν οφέλη για την υγεία.

1.5. Τοξικότητα των νανοϋλικών

Το εξαιρετικά μικρό μέγεθος των νανοϋλικών σε συνδυασμό με τις σπάνιες ιδιότητες που είναι απόρροια των φυσικοχημικών χαρακτηριστικών τους εκτός από τις αναρίθμητες δυνατότητες αξιοποίησής τους έχει δημιουργήσει και έντονη ανησυχία για τους κινδύνους που μπορεί να ελλοχεύουν από την τοξικότητά τους. Τα νανοϋλικά

μπορούν να εισέλθουν στον οργανισμό μέσω της αναπνοής ή της τροφικής αλυσίδας και το μικρό τους μέγεθος παρέχει την δυνατότητα διάσχισης κυτταρικών μεμβρανών και μετακίνησης κατά μήκους των αξόνων και των δενδριτών που συνδέουν τους νευρώνες. Το αντίκτυπο της τοξικότητας μπορεί να είναι σαφώς αρνητικό αλλά μπορεί να είναι και θετικό αφού η κατανόηση και η ελεγχόμενη χρήση της θα μπορούσε να συμβάλει στην αξιοποίηση των νανοϋλικών στην ιατρική εξασφαλίζοντας όλα τα οφέλη που αναπτύχθηκαν στην παράγραφο 1.4.1.

2. Μηχανική Μάθηση

2.1. Γενικά

Τα τελευταία χρόνια ως συνέπεια της ανάπτυξης στον τομέα της επιστήμης των ηλεκτρονικών υπολογιστών παρουσιάζει ιδιαίτερο ενδιαφέρον η μηχανική μάθηση (machine learning). Πρόκειται για έναν κλάδο της προαναφερθείσας επιστήμης που σκοπό έχει την ένταξη της υπολογιστικής θεωρίας μάθησης και των προτύπων στην τεχνητή νοημοσύνη. Ο πρώτος ορισμός της μηχανικής μάθησης μας γυρίζει αρκετά χρόνια πίσω, το 1959, όταν ο Αμερικανός Arthur Samuel, ο οποίος θεωρείται πρωτοπόρος της Τεχνητής Νοημοσύνης, την όρισε ως το *"Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί"*.[24] Τον ορισμό του Samuel ακολούθησαν και άλλοι οι οποίοι έκαναν προσπάθειες να ορίσουν την μηχανική μάθηση. Το 1987 ο Carbonell, διατύπωσε την άποψη πως η μηχανική μάθηση είναι «... η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης», δέκα χρόνια ο αργότερα, το 1997 ο Mitchell υποστήριζε πως *«Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μία μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E»* για να έρθουν στο κατώφλι του 21ου αιώνα οι Witten & Frank να υποστηρίξουν πως *«Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον»*.[35] Θα μπορούσαμε κοινώς να πούμε πως στη μηχανική μάθηση αντί ο χρήστης (προγραμματιστής) να προγραμματίσει τον ηλεκτρονικό υπολογιστή, κάνει χρήση αλγορίθμων και πειραματικών δεδομένων κατασκευάζοντας μοντέλα, ώστε ο υπολογιστής να «μάθει» μόνος του και να οδηγηθεί σε προβλέψεις ή αποφάσεις που είναι και το αποτέλεσμα, προσομοιάζοντας κατά κάποιο τρόπο την φυσική, ανθρώπινη μάθηση.

Με δεδομένο λοιπόν αυτό και λαμβάνοντας υπόψη πως η έννοια της μάθησης σε ένα σύστημα σχετίζεται με την ικανότητά του συστήματος μέσω της αλληλεπίδρασής του με το περιβάλλον δραστηριοποίησής του να αποκτά επιπλέον γνώσεις αλλά και την ικανότητα βελτίωσης εκτέλεσης μιας ενέργειας ή διεργασίας μέσω των επαναλήψεων μπορεί να γίνει εύκολα αντιληπτό ότι τα συστήματα με ικανότητα μηχανικής μάθησης

είναι «έξυπνα συστήματα» με αυξημένες δυνατότητες όπως αυτή της συνεχούς βελτίωσης σε σχέση με τις λειτουργίες που μπορούν να επιτελέσουν, της μεταβολής της γνωσιακής τους βάσης είτε μετασχηματίζοντας την εσωτερική τους δομή, π.χ. όταν πρόκειται για νευρωνικά δίκτυα, είτε αποκτώντας πρόσβαση σε επιπλέον γνώση, όπως κάνουν τα έμπειρα συστήματα και τέλος της δυνατότητα γενίκευσης μέσω της οποία αγνοούν ιδιότητες και χαρακτηριστικά με μικρή συμβολή στην έννοια-ενέργεια που πρέπει να μάθουν.[48]

2.2. Σχέση με την στατιστική

Οι τομείς της μηχανικής μάθησης και της στατιστικής φαίνεται να είναι άρρηκτα συνδεδεμένοι καθώς όλο και περισσότεροι στατιστικοί κάνουν χρήση μεθόδων μηχανικής μάθησης με αποτέλεσμα τη δημιουργία ενός ακόμη πεδίου, αυτού της στατιστικής μάθησης. Αναφορικά με την σχέση μηχανικής μάθησης και στατιστικής έχουν γίνει πολλές τοποθετήσεις με χαρακτηριστικότερες αυτές του Καναδού Στατιστικού Larry Wasserman, πως «*Και τα δύο ασχολούνται με το ίδιο ερώτημα: πώς μπορούμε να μάθουμε από τα δεδομένα*;» και του Αμερικανού Επιστήμονα Michael I. Jordan, σύμφωνα με τον οποίο οι βασικές αρχές και εργαλεία που χρησιμοποιούνται στην μηχανική μάθηση προϋπήρχαν και έχουν δανειστεί από την στατιστική, ο οποίος πρότεινε μάλιστα οι δύο αυτοί τομείς να ενωθούν υπό την σκέπη ενός κοινού πεδίου για τον οποίο πρότεινε τον ορισμό Επιστήμη Δεδομένων.[18] Σε συνέχεια των προηγουμένων ο Αμερικανός Στατιστικός Leo Breiman παρατήρησε την ταύτιση του αλγοριθμικού μοντέλου της στατιστικής μοντελοποίησης με αλγόριθμους μηχανικης μαθησης όπως τα Τυχαία Δάση (Random Forests).[54]

Παρά τις σημαντικές τους ομοιότητες και το γεγονός ότι έχουν κοινό στόχο, η μηχανική μάθηση και η στατιστική έχουν διαφορετική βάση καθώς η μεν στατιστική αποτελεί πεδίο των μαθηματικών και υπήρχε πολύ πριν την εφεύρεση των ηλεκτρονικών υπολογιστών, ενώ η μηχανική μάθηση είναι ένα νέο επιστημονικό πεδίο που προέρχεται από την τεχνητή νοημοσύνη και η άνθισή του είναι συνάρτηση της ανάπτυξης της

επιστήμης των υπολογιστών. Σημαντική διαφορά αποτελεί επίσης η επεξεργασία των δεδομένων αφού σε αντίθεση με τη στατιστική που απαιτεί κατανόηση της συλλογής των δεδομένων και επιλογή των κατάλληλων παραμέτρων για την επιτυχή πρόβλεψη, η μηχανική μάθηση αγνοεί οποιαδήποτε σχέση μεταξύ των μεταβλητών αφού οι αλγόριθμοι χρησιμοποιούν όλα τα δεδομένα και καταλήγουν στην επιλογή των παραμέτρων που θα οδηγήσουν σε μια επιτυχημένη πρόβλεψη και μάλιστα όσο αυξάνεται ο αριθμός των διαθέσιμων δεδομένων, αυξάνεται και η ακρίβεια της πρόβλεψης. Γι' αυτόν ακριβώς το λόγο η στατιστική χρησιμοποιεί μικρού όγκου δεδομένα, σε αντίθεση με τη μηχανική μάθηση που βρίσκει εφαρμογές σε περιπτώσεις όπου είναι διαθέσιμες μεγάλες δεξαμενές δεδομένων. Η βασικότερη όμως ίσως διαφορά μεταξύ των δύο πεδίων έγκειται στην προσέγγιση του ζητήματος που καλούνται να αντιμετωπίσουν καθώς στόχος της μηχανικής μάθησης είναι η βελτιστοποίηση και η αύξηση της αποτελεσματικότητας ενώ η στατιστική εστιάζει στο συμπέρασμα αυτό κάθε αυτό. Η διαφορά αυτή γίνεται πιο κατανοητή από το παράδειγμα που παρατίθεται παρακάτω και το οποίο αποτυπώνει την περιγραφή του αποτελέσματος του ίδιου μοντέλου από έναν στατιστικό και έναν Machine Learning Engineer.[51]

Machine Learning Engineer	«Το μοντέλο είναι 85% ακριβές στην πρόβλεψη του Υ δεδομένου των α, β και γ»
Στατιστικός	«Το μοντέλο είναι 85% ακριβές στην πρόβλεψη του Υ δεδομένου των α, β και γ και είμαι 90% σίγουρος ότι αν ξανακάνεις το πείραμα θα επιτευχθεί το ίδιο αποτέλεσμα»

2.3. Σχέση με άλλους τομείς

Εκτός από την στατιστική με την οποία η μηχανική μάθηση είναι στενά συνδεδεμένες, ο τομέας της μηχανικής μάθησης έχει στενούς δεσμούς με τους τομείς της τεχνητης νοημοσύνης (Artificial Intelligence), της εξόρυξης δεδομένων (Data Mining) αλλά και της βελτιστοποίησης (optimization).

2.3.1. Τεχνητή Νοημοσύνη

Ήδη έχει αναφερθεί πως η μηχανική μάθηση εμφανίστηκε ως εξέλιξη της τεχνητής νοημοσύνης και ευνοήθηκε ιδιαίτερα από την ψηφιακή διαθεσιμότητα δεδομένων και της δυνατότητας αυτά να διανεμηθούν διαδικτυακά, ωστόσο κρίνοντας από το γεγονός πως οι δύο αυτές φράσεις χρησιμοποιούνται συχνά η μια ως εναλλακτικής της άλλης φαίνεται πως η διαφορά τους δεν είναι ξεκάθαρη. Είναι αλήθεια πως όταν το θέμα είναι ο μεγάλος όγκος δεδομένων (big data) γίνεται χρήση και των δύο όρων και αυτό οδηγεί συχνά σε σύγχυση. Για τον λόγο αυτό θα ήταν χρήσιμο να ξεκαθαρίσουμε πως η τεχνητή νοημοσύνη αποτελεί το ευρύτερο πεδίο έρευνας που σκοπό έχει την «εκπαίδευση» των μηχανών ώστε να εκτελούν τις εργασίες που τους αναθέτουμε με «έξυπνο» τρόπο ενώ η μηχανική μάθηση είναι μια εφαρμογή των αρχών της τεχνητής νοημοσύνης που βασίζεται στην ιδέα ότι είμαστε σε θέση να παρέχουμε στους υπολογιστές πρόσβαση στα δεδομένα και να τους αφήσουμε να εκπαιδευτούν μόνοι τους πάνω στον τρόπο που θα εκτελέσουν τις εργασίες που εμείς θέλουμε.

2.3.2. Εξόρυξη δεδομένων (Data Mining)

Όσον αφορά την εξόρυξη δεδομένων και την σχέση της με την μηχανική μάθηση συχνά συγχέονται πιθανότατα λόγω του γεγονότος ότι συχνά χρησιμοποιούν τις ίδιες μεθόδους. Παρ' όλα αυτά διαφέρουν σημαντικά αφού η εξόρυξη δεδομένων εστιάζει στην ανακάλυψη ιδιοτήτων και πληροφοριών που εξάγονται από μια μεγάλη δεξαμενή δεδομένων και χρησιμοποιούνται για την δημιουργία εφαρμογών ικανών να αξιοποιήσουν τις πληροφορίες που ανακαλύπτονται. Ακολούθως η μηχανική μάθηση χρησιμοποιεί τα δεδομένα που εξάγονται από τις εξορύξεις και με τη χρήση αλγορίθμων δημιουργούνται οι επιθυμητές ενέργειες, εστιάζει συνεπώς στην πρόβλεψη, που βασίζεται σε γνωστές ιδιότητες που απορρέουν από το σύνολο εκπαίδευσης. Θα μπορούσαμε λοιπόν να παρομοιάσουμε την εξόρυξη δεδομένων ως το καύσιμο που τροφοδοτεί τον κινητήρα της

μηχανικής μάθησης και κρίνεται απαραίτητο για τη λειτουργία του, όπως απαραίτητα είναι και τα δεδομένα που προκύπτουν από την εξόρυξη για τους αλγόριθμους της μηχανικής μάθησης.

2.3.3. Βελτιστοποίηση (optimization)

Στενή είναι τέλος η σχέση μεταξύ της μηχανικής μάθησης και της βελτιστοποίησης, της διαδικασίας δηλαδή κατά την οποία επιχειρείται η ελαχιστοποίηση της διαφοράς μεταξύ της πρόβλεψης και της πραγματικότητας για το υπο μελέτη πρόβλημα που επιτυγχάνεται με την ελαχιστοποίηση της συνάρτησης απώλειας (loss function). Στο μεγαλύτερο ποσοστό τους τα προβλήματα μηχανικής μάθησης καταλήγουν να είναι προβλήματα βελτιστοποίησης. [15]

2.4. Είδη Μηχανικής Μάθησης

Σε έναν τομέα ταχύτατα αναπτυσσόμενο όπως είναι αυτός της μηχανικής μάθησης αντιλαμβανόμαστε ότι αναπτύσσονται διαρκώς νέες τεχνικές οι οποίες ανήκουν σε ένα από τα παρακάτω είδη. Της επιβλεπόμενης μάθησης (supervised learning), της μη επιβλεπόμενης μάθησης (unsupervised learning) ή της ενισχυτικής μάθησης (reinforcement learning).

2.4.1. Επιβλεπόμενη μάθηση (supervised learning)

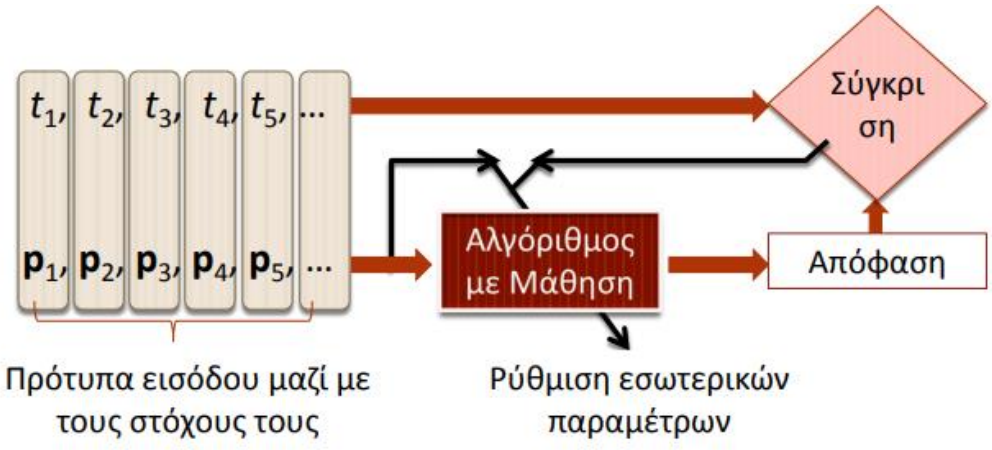
Στην επιβλεπόμενη μάθηση ο στόχος είναι η εκπαίδευση σε έναν κανόνα ή αλλιώς η δημιουργία μιας συνάρτησης με τη χρήση πραγματικών δεδομένων εισόδου και επιθυμητών δεδομένων εξόδου (αποτελεσμάτων) και η γενίκευσή της ώστε να βρίσκει εφαρμογή και σε περιπτώσεις που τα αποτελέσματα είναι άγνωστα. Η συνάρτηση αυτή

ονομάζεται συνάρτηση στόχος (target function) χρησιμοποιείται για την πρόβλεψη της τιμής της μεταβλητή εξόδου (εξαρτημένη μεταβλητή), βάσει των τιμών των μεταβλητών εισόδου (ανεξάρτητες μεταβλητές) και η όλη ιδέα της επιβλεπόμενης μάθησης βασίζεται υπόθεση επαγωγικής μάθησης (inductive learning hypothesis), σύμφωνα με την οποία:

«Κάθε υπόθεση h που προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν έχει εξετάσει».[35]

Η επιβλεπόμενη μάθηση χρησιμοποιείται στις ακόλουθες περιπτώσεις προβλημάτων:

- Τα προβλήματα παλινδρόμησης (**regression**) που αφορούν την δημιουργία μοντέλων που σκοπό έχουν την πρόβλεψη αριθμητικών τιμών.
- Τα προβλήματα ταξινόμησης (**classification**) που στοχεύουν στην δημιουργία μοντέλων πρόβλεψης διακριτών κατηγοριών-τάξεων.



Εικόνα 2: Επιβλεπόμενη Μάθηση
Πηγή: Κ. Διαμαντάρας, Μηχανική Μάθηση – Μάθημα 1, Βασικές Έννοιες, Τμήμα Πληροφορικής, ΤΕΙ Θεσσαλονίκης, 2011

2.4.2. Μη επιβλεπόμενη μάθηση (unsupervised learning)

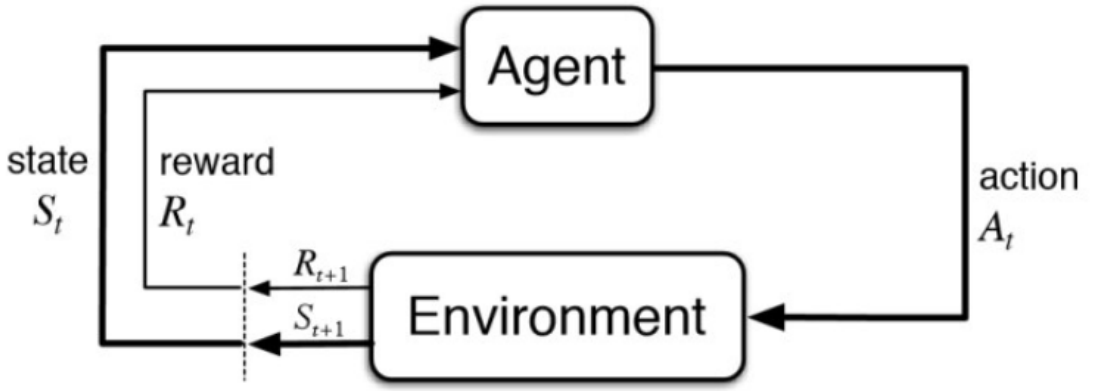
Στην μη επιβλεπόμενη μάθηση ο αλγόριθμος πρέπει να ανακαλύψει την δομή των δεδομένων εισόδου και στη συνέχεια να κατασκευάσει μοντέλα συσχέτισης για κάποιο σύνολο δεδομένων υπό μορφή παρατηρήσεων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Χρήση μεθόδων μη επιβλεπόμενης μάθησης γίνεται κυρίως σε προβλήματα ανάλυσης συσχετισμών (Association Analysis) όπου στόχος είναι η ανακάλυψη σχέσεων σε μεγάλα σύνολα δεδομένων αλλά και ομαδοποίησης (Clustering)[36] όπου από ένα σύνολο δεδομένων δημιουργούνται ομάδες με τέτοιο τρόπο ώστε τα αντικείμενα της ίδιας ομάδας (που ονομάζεται σύμπλεγμα) να έχουν μεταξύ τους περισσότερες ομοιότητες σε σχεση με τα υπόλοιπα που ανήκουν σε άλλες ομάδες (ομάδες). Να σημειωθεί ότι το clustering αποτελεί στάδιο της διερευνητικής εξόρυξης δεδομένων που αναφέρθηκε νωρίτερα.

Εικόνα 4: Μη Επιβλεπόμενη Μάθηση
Πηγή: Κ. Διαμαντάρας, Μηχανική Μάθηση – Μάθημα 1, Βασικές Έννοιες, Τμήμα Πληροφορικής, ΤΕΙ Θεσσαλονίκης, 2011

2.4.3. Ενισχυτική μάθηση (reinforcement learning)

Στην ενισχυτική μάθηση ο αλγόριθμος εκπαιδεύεται σε μια στρατηγική ενεργειών μέσω της αλληλεπίδρασης του με ένα διαδραστικό περιβάλλον μέσω δοκιμών και σφαλμάτων και χρησιμοποιώντας ανατροφοδότηση από τις δικές του ενέργειες και

εμπειρίες. Σε αντίθεση με την επιβλεπόμενη μάθηση όπου η ανατροφοδότηση που παρέχεται στον αλγόριθμο είναι το σωστό σύνολο ενεργειών για την εκτέλεση μιας εργασίας, η ενισχυτική μάθηση χρησιμοποιεί ανταμοιβές και τιμωρίες ως σήματα θετικής και αρνητικής συμπεριφοράς και σκοπός της είναι να μεγιστοποιήσει μια συνάρτηση του αριθμητικού σήματος ενίσχυσης (ανταμοιβή), για παράδειγμα την αναμενόμενη τιμή του σήματος ενίσχυσης στο επόμενο βήμα χωρίς να παρέχεται κάποια ενημέρωση από κάποιον εξωτερικό επιβλέποντα για το κατά πόσο βρίσκεται κοντά στην επίτευξη του στόχου. Αυτή η κατηγορία είναι ιδιαιτέρως χρήσιμη στις περιπτώσεις που καλούμαστε να αντιμετωπίσουμε προβλήματα σχεδιασμού (Planning), όπως για παράδειγμα ο έλεγχος κίνησης ρομπότ και η βελτιστοποίηση εργασιών σε εργοστασιακούς χώρους.



Εικόνα 5: Ενισχυτική Μάθηση
Πηγή: https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html

2.5. Αλγόριθμοι Μηχανικής Μάθησης

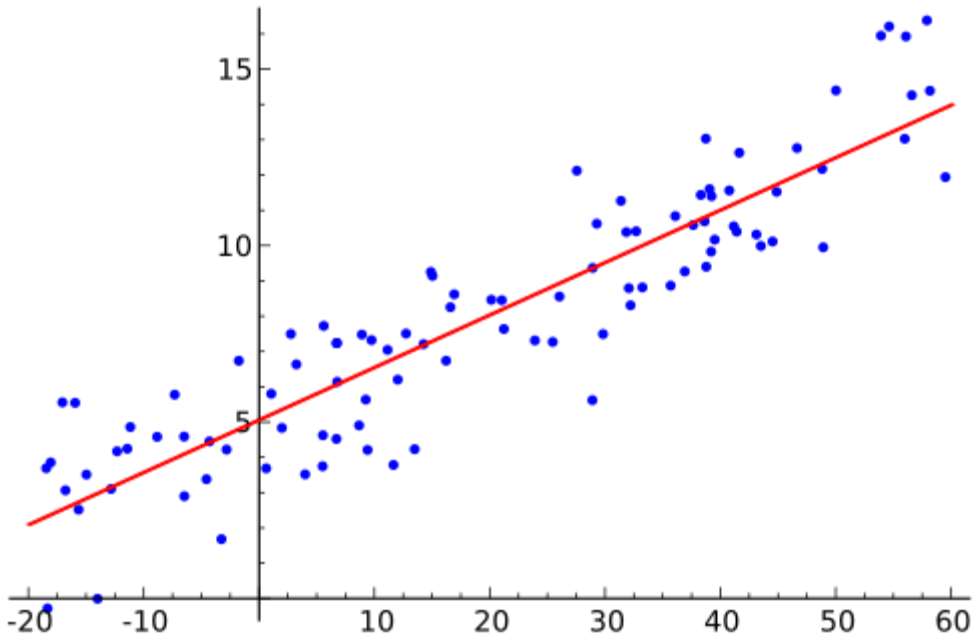
2.5.1. Γραμμική Παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι μια προσέγγιση που στοχεύει στην δημιουργία, την περιγραφή και την αξιολόγηση των σχέσεων μεταξύ μιας μεταβλητής (y) που θεωρείται τυχαία και ονομάζεται εξαρτημένη και μιας ή περισσοτέρων μη τυχαίων

μεταβλητών (x1, x2, x3……..) που καλούνται ανεξάρτητες. Στόχος της είναι να εξετάσει πως μια μεταβολή στις ανεξάρτητες μεταβλητές επηρεάζει την τιμή της εξαρτημένης μεταβλητής. Στην περίπτωση που υπάρχει μόνο μια ανεξάρτητη μεταβλητή μιλάμε για απλή γραμμική παλινδρόμηση (Simple Linear Regression) ενώ όταν οι ανεξάρτητες μεταβλητές είναι περισσότερες από μια το είδος της γραμμικής παλινδρόμησης ονομάζεται πολλαπλή γραμμική παλινδρόμηση (Multiple Linear Regression).[44]

Η γραμμική παλινδρόμηση εκτελείται σε δύο φάσεις, στην Α' Φάση εξετάζεται η ύπαρξη της σχέσης μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής και αν βρεθεί σχέση που συνδέει τις δύο αυτές μεταβλητές υλοποιείται η Β' Φάση κατά την οποία πραγματοποιείται η ανάλυση παλινδρόμησης και η μοντελοποίηση των δεδομένων με γραμμικές λειτουργίες που οδηγεί στην δημιουργία του γραμμικού μοντέλου. Η μορφή ενός απλού γραμμικού μοντέλου, με τον όρο e_i να συμβολίζει το τυχαίο σφάλμα ή αλλιώς τον θόρυβο με μηδενική μέση τιμή και διασπορά σ^2, είναι η ακόλουθη:

$$f(x_i) = y_i = a + \beta x_i + e_i$$



Εικόνα 6: Παράδειγμα απλής γραμμικής παλινδρόμησης. Διάγραμμα διασποράς τιμών {x, y} με ανεξάρτητη μεταβλητή την x. Η κόκκινη ευθεία είναι η βέλτιστη εξίσωση y=α+βx που μοντελοποιεί τα σημεία
Πηγή: https://el.wikipedia.org/

Έχοντας λοιπόν στην διάθεσή μας το μοντέλο, είμαστε σε θέση να προβλέψουμε την τιμή της εξαρτημένης μεταβλητής y για κάθε τιμή του x ενώ παρέχεται και η δυνατότητα ποσοτικοποίησης της αντοχής της σχέσης μεταξύ y και x, ώστε να αξιολογηθεί η μεταξύ τους σχέση και να εντοπιστεί ποιες υποκατηγορίες του x περιέχουν περιττές πληροφορίες σχετικά με το y. [17] Δεδομένου ότι ο στόχος της γραμμικής παλινδρόμησης είναι η εύρεση της κατάλληλης συνάρτησης που θα αποτυπώνει την σχέση μεταχύ των μεταβλητών $\{x_i, y_i\}$ αντιλαμβανόμαστε τη σπουδαιότητα ορθής εκτίμησης των παραμέτρων α και β.

Η πιο διαδεδομένη μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων α και β και συνεπώς για την εύρεση της εξίσωσης της ευθείας που μοντελοποιεί καλύτερα τα δεδομένα, είναι η μέθοδος ελαχίστων τετραγώνων (least squares) που χρησιμοποιήθηκε για πρώτη φορά το 1805 από τον Γάλλο μαθηματικό Legendre και στοχεύει στην ελαχιστοποίηση του αθροίσματος των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (xi, yi) από την ευθεία $y_i=\alpha+\beta x_i$ δηλαδή την ελαχιστοποίηση του $\sum_{i=1}^{v} e_i^{\,2} = \sum_{i=1}^{v} (y_i - a + \beta x_i)^2$. Οι τιμές α και β που ελαχιστοποιούν την παραπάνω εξίσωση ονομάζονται εκτιμήτριες ελαχίστων τετραγώνων (least square estimators) και υπολογίζονται από τις παρακάτω σχέσεις: [42]

$$\hat{\beta} = \frac{v\sum_{i=1}^{v} x_i y_i - \left(\sum_{i=1}^{v} x_i\right)\left(\sum_{i=1}^{v} y_i\right)}{v\sum_{i=1}^{v} x_i^2 - \left(\sum_{i=1}^{v} x_i\right)^2} \qquad \text{όπου } \bar{y} = \frac{1}{v}\sum_{i=1}^{v} y_i, \quad \bar{x} = \frac{1}{v}\sum_{i=1}^{v} x_i .$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Ενώ η ευθεία $\hat{y} = \hat{\alpha} + \hat{\beta}x$ καλείται ευθεία ελαχίστων τετραγώνων.

Στην μηχανική μάθηση εκτός από την μέθοδο ελαχιστων τετραγώνων συχνά για την εκτίμηση των παραμέτρων $\min_{\theta_0,\theta_1} J(\theta_0,\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$ χρησιμοποιείται ο αλγόριθμος απότομης καθόδου (Gradient descent). Έχοντας λοιπόν μια συνάρτηση γραμμικής παλινδρόμησης που ονομάζεται συνάρτηση υπόθεσης $h_\theta(x) = \theta_0 + \theta_1 x$ στόχος μας είναι η ελαχιστοποίηση της συνάρτησης κόστους ελάχιστων τετραγώνων δηλαδή

όπου m ο αριθμός των δειγμάτων {x,y}. Ο αλγόριθμος της απότομης καθόδου εκκινεί ορίζοντας αρχικές τιμές στα θ_0 και θ_1 και τις μεταβάλλει διαρκώς μέχρι να συγκλίνει στις τιμές εκείνες που ελαχιστοποιούν την τιμή της εξίσωσης $J(\theta_0, \theta_1)$. [2]

Τα παραπάνω ισχύουν στην περίπτωση που εξαρτημένη μεταβλητή είναι γραμμική συνάρτηση μιας μόνο ανεξάρτητης μεταβλητής, όταν όμως αποτελεί γραμμικό συνδυασμό m ανεξάρτητων μεταβλητών, όπως συμβαίνει στην πλειονότητα των περιπτώσεων η σχέση $y_i = a + \beta x_i + e_i$ διαμορφώνεται ως εξής: $y_i = a + b_1 x_{1i} + b_2 x_2 + b_3 x_3 + ... + b_m x_m + e$.

όπου:

y_i : τιμή της εξαρτημένης μεταβλητής

$x_1, x_2, x_3, \ldots x_m$: τιμές των ανεξάρτητων μεταβλητών

a: σταθερά

$b_1, b_2, b_3 \ldots, b_m$: συντελεστές παλινδρόμησης ενδεικτικοί της επίδρασης μεταβλητών x_i

e : σφάλμα

Αφού λοιπόν η εξαρτημένη μεταβλητή y είναι συνάρτηση m ανεξάρτητων μεταβλητών, η μεταβλητότητά της επηρεάζεται από την μεταβλητότητα παραπάνω από μιας μεταβλητής, σε αντίθεση με την απλή παλινδρόμηση. Στην πολλαπλή γραμμική παλινδρόμηση η τιμή της εξαρτημένης μεταβλητής y αποτελεί συνάρτηση δύο συνιστωσών, του μη τυχαίου παράγοντα που ενσωματώνει τις συστηματικές επιδράσεις των $x_1, x_2, x_3, \ldots x_m$, και του τυχαίου παράγοντα του σφάλματος e , που ενσωματώνει όλους τους άλλους (εκτός των $x_1, x_2, x_3, \ldots x_m$) παράγοντες που επηρεάζουν την τιμή της της εξαρτημένης μεταβλητής y.

Βασικό στάδιο στην δημιουργία μοντέλων γραμμικής παλινδρόμησης είναι η υιοθέτηση κάποιων παραδοχών με τις συνηθέστερες να είναι η γραμμικότητα, η υπόθεση δηλαδή πως η μέση τιμή της μεταβλητής απόκρισης αποτελεί γραμμικό συνδυασμό των συντελεστών παλινδρόμησης και των μεταβλητών πρόβλεψης και η ανεξαρτησία των λαθών, η υπόθεση δηλαδή πως τα λάθη των μεταβλητών απόκρισης δεν έχουν κάποια σχέση μεταξύ τους και κατανέμονται κανονικά, επίσης θεωρείται πως οι αναμενόμενες τιμές (μέσοι) των σφαλμάτων e_i είναι μηδέν και τα σφάλματα e_i έχουν την ίδια διακύμανση $\sigma^2 e$ για όλους τους συνδυασμούς των τιμών των ανεξάρτητων μεταβλητών. Τέλος, στην πολλαπλή παλινδρόμηση καλό είναι οι ανεξάρτητες μεταβλητές να είναι ανεξάρτητες μεταξύ τους, και στην περίπτωση ύπαρξης συσχέτισης μεταξύ των μεταβλητών θα πρέπει να επιλέγεται μόνο μια εξ' αυτών έτσι ώστε να αποφεύγεται η δημιουργία του προβλήματος της πολυσυγγραμικότητας[41] ενώ ο πίνακας που περιλαμβάνει τις ανεξάρτητες μεταβλητές Χ που λαμβάνουν μέρος στην ανάλυση καλό είναι να περιέχει λίγες στήλες και πολλές γραμμές, δηλαδή πολλά αντικείμενα και λίγες μεταβλητές.

Η γραμμική παλινδρόμηση ήταν ο πρώτος τύπος της ανάλυσης παλινδρόμησης που μελετήθηκε και χρησιμοποιήθηκε εκτενέστατα σε πληθώρα εφαρμογών λόγω της ευκολίας των γραμμικών μοντέλων αλλά και της ευκολίας προσδιορισμού των στατιστικών ιδιοτήτων των προκυπτόντων εκτιμήσεων.

2.5.2. Πολυμεταβλητή Ανάλυση Δεδομένων (MultiVariate Data Analysis, MVDA)

Πρόκειται για ένα είδος ανάλυσης δεδομένων το οποίο παρέχει τις δυνατότητες ανάλυσης πάνω από μιας μεταβλητών απόκρισης και ταυτόχρονα της διαχείρισης πολλών συσχετιζόμενων ανεξάρτητων μεταβλητών. Η πολυμεταβλητή ανάλυση δεδομένων ακολουθεί την μέθοδο της προβολής των σημείων από ένα πολυδιάστατο χώρο σε ένα χώρο μικρότερων διαστάσεων, ενώ δεν ακολουθεί την τακτική της αλλαγής μία παραμέτρου τη φορά.[33] Κάτω από την ομπρέλα της πολυμεταβλητής ανάλυσης δεδομένων

(MVDA) βρίσκονται, μεταξύ άλλων, οι μέθοδοι της Ανάλυσης Κυρίων Συνιστωσών (Principal Component Analysis, PCA) και της Ανάλυσης Μερικών Ελαχίστων Τετραγώνων (Partial Least Squares, PLS) που παρουσιάζονται στις αμέσως επόμενες παραγράφους.

2.5.2.1. Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis)

Η ανάλυση κύριων συνιστωσών είναι μια στατιστική διαδικασία που χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό για να μετατρέψει ένα συνόλου παρατηρήσεων πιθανών συσχετισμένων μεταβλητών (οντότητες που λαμβάνουν διάφορες αριθμητικές τιμές) σε ένα σύνολο τιμών οι οποίες έχουν την ιδιότητα να είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και παράλληλα να μη συσχετίζονται μεταξύ τους γραμμικά και ονομάζονται κύρια συστατικά. Αποτελεί μια απλή μέθοδο πολυμεταβλητής ανάλυση και στοχεύει στην ανεύρεση διακριτών ολιγάριθμων μεταβλητών από ένα πλήθος p μεταβλητών. Έτσι αν για παράδειγμα υπάρχουν διαθέσιμες n παρατηρήσεις με πλήθος μεταβλητών p, τότε ο αριθμός των διακριτών κύριων συνιστωσών είναι θα είναι min (n-1, p). Ο μετασχηματισμός αυτός ορίζεται με τέτοιο τρόπο ώστε το πρώτο κύριο συστατικό να έχει τη μεγαλύτερη δυνατή διακύμανση, αποδίδωντας έτσι τη μεγαλύτερη μεταβλητότητα στα δεδομένα, και κάθε επόμενο στοιχείο με τη σειρά του έχει τη μεγαλύτερη δυνατή διακύμανση. Το μεγάλο πλεονέκτημά τους έγκειται στην ιδιαιτερότητα που διαθέτουν, λόγω της ανάλυσης, να εξηγούν πολύ μεγάλο ποσοστό της ολικής μεταβλητότητας που αναπτύσσεται μεταξύ των p μεταβλητών, το οποίο τελικά κατανέμεται σε μερικές μόνο νέες μεταβλητές.

Βασικούς άξονες της ανάλυσης κύριων συνιστωσών αποτελούν αφενός το γεγονός ότι από το p πλήθος μεταβλητών x_1, x_2,…..x_p παράγονται ισάριθμοι συνδυασμοί των μεταβλητών, οι z_1, z_2,…..z_p, χωρίς να σχετίζονται μεταξύ τους κάτι που προδιαθέτει την ικανότητα μέτρησης διαφορετικών διαστάσεων των στοιχείων και αφετέρου το γεγονός ότι οι διακυμάνσεις (μεταβλητότητα) που αναπτύσσονται μεταξύ των μεταβλητών zi, διαβαθμίζονται με τέτοιο τρόπο ώστε η πρώτη μεταβλητή z_1 να είναι σε θέση να επεξηγεί

ένα όσο το δυνατόν μέγιστο ποσοστό της ολικής μεταβλητότητας, η z_2 ένα δεύτερο μέγιστο ποσοστό αυτής κ.ο.κ., υπακούοντας στη σχέση: $\lambda_1>\lambda_2>...>\lambda_p$, όπου λi η i ποσότητα της διακύμανσης. Η τεχνική των κύριων συνιστωσών έχει ως βάση, κατά τη διαδικασία υπολογισμού της, τη μήτρα των κατά ζεύγη συσχετίσεων (correlation matrix) των μεταβλητών. Κατά συνέπεια, για να θεωρείται η τεχνική επιτυχημένη, να παρέχει δηλαδή ουσιώδη πληροφόρηση, απαραίτητη προϋπόθεση είναι κάποιοι συντελεστές συσχέτισης των αρχικών μεταβλητών της μήτρας συσχετίσεων να φέρουν υψηλές τιμές θετικές ή αρνητικές (π.χ. r≥±0,700) διατηρώντας όμως μια ισορροπία αφού αρχικές μεταβλητές με πολύ ισχυρές τιμές συσχετίσεων (π.χ. r>±0,990) θεωρούνται πλεονάζουσες και συνίσταται να μην συμπεριλαμβάνονται στην εφαρμογή της μεθόδου.

Το πρώτο βήμα για την ανάλυση κυρίων συνιστωσών είναι ο μετασχηματισμός των μεταβλητών σύμφωνα με την σχέση: $\frac{(Xi-\overline{X})}{s}$. Στην συνέχεια από τον γραμμικό συνδυασμό των p μεταβλητών παράγεται η πρώτη κύρια συνιστώσα $Z_1=\alpha_{11}X_1+\alpha_{12}X_2+.....+\alpha_{1p}X_p$ και ακολουθούν οι υπόλοιπες οι οποίες προκύπτουν με τον ίδιο τρόπο (π.χ. $Z_2=\alpha_{21}X_1+\alpha_{22}X_2+.....+\alpha_{2p}X_p$, $Z_3=\alpha_{31}X_1+\alpha_{32}X_2+.....+\alpha_{3p}X_p$) εως ότου δημιουργηθούν τόσες συνιστώσες (p) συνιστώσες Zi όσες είναι και οι αρχικές μεταβλητές. Να σημειωθεί ότι βασικός περιορισμός είναι ο συντελεστής συσχέτισης των συνιστωσών να ισούται με μηδέν, να είναι δηλαδή ασυσχέτιστες μεταξύ τους. Ο συντελεστής α_{ij} που εμφανίζεται στις εξισώσεις των συνιστωσών ονομάζεται ειδικός συντελεστής στάθμισης της j μεταβλητής στην i συνιστώσα ο οποίος για να εξασφαλίσει την εκτίμηση της μέγιστης διακύμανσης των μεταβλητών z διέπεται από τον περιορισμό $\alpha_{11}^2+\alpha_{12}^2+.....+\alpha_{1p}^2=1$ για την πρώτη συνιστώσα, $\alpha_{21}^2+\alpha_{22}^2+.....+\alpha_{2p}^2=1$ για την δεύτερη συνιστώσα, $\alpha_{31}^2+\alpha_{32}^2+.....+\alpha_{3p}^2=1$ για την τρίτη συνιστώσα κ.ο.κ.[47] Οι συντελεστές στάθμισης aij υπολογίζονται με τη βοήθεια της μήτρας C των συνδιακυμάνσεων των αρχικών μεταβλητών,

$$C = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ c_{p1} & c_{p2} & \dots & c_{pp} \\ \end{pmatrix}$$ όπου τα διαγώνια στοιχεία cii είναι οι διακυμάνσεις της Xi

και cij οι συνδιακυμάνσεις των μεταβλητών Xi και Xj

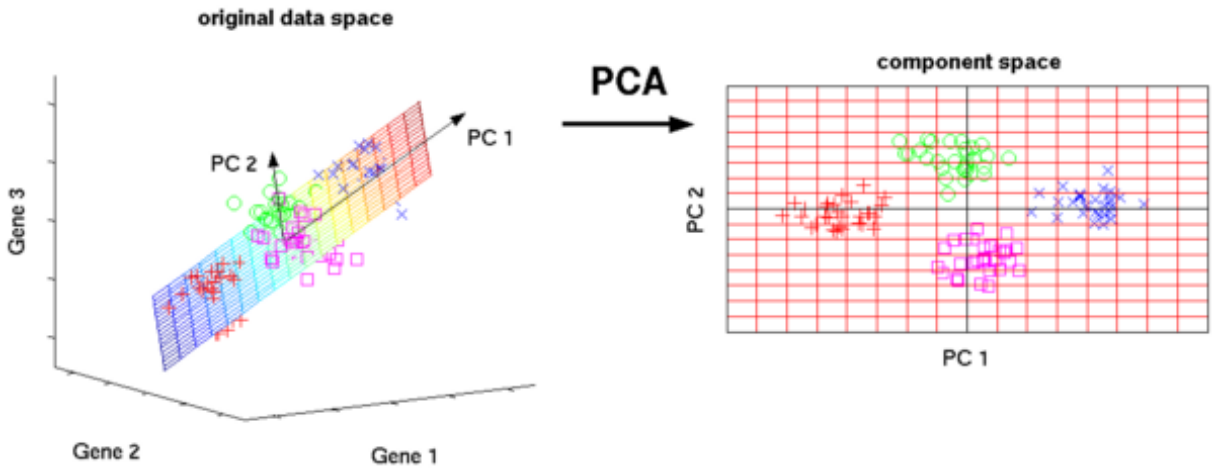
Με την τυποποίηση των αρχικών μεταβλητών η μήτρα των συνδιακυμάνσεων μεταπίπτει στη μήτρα των συσχετίσεων που είναι και το βασικό στοιχείο της ανάλυσης κύριων συνιστωσών.

$$C = \begin{pmatrix} 1 & c_{12} & \dots & c_{1p} \\ c_{21} & 1 & \dots & c_{2p} \\ c_{p1} & c_{p2} & \dots & 1 \\ \end{pmatrix}$$ έτσι προκύπτει cii=1, ενώ cij=cji είναι ο συντελεστής συσχέτισης μεταξύ των Xi και Xj

. Οι διακυμάνσεις των κύριων συνιστωσών λi ονομάζονται ιδιοτιμές ή χαρακτηριστικές ρίζες και για αυτές ισχύει λ1,λ2,...,λP > 0 και λ1+λ2+...λP= c11+c22+...+cpp το άθροισμά τους δηλαδή ισοδυναμεί με το άθροισμα των διακυμάνσεων των αρχικών μεταβλητών. Οι συσχετίσεις rij μεταξύ των αρχικών μεταβλητών και των κύριων συνιστωσών ονομάζονται φορτία και δείχνουν την ένταση της δράσης που αναπτύσσουν οι αρχικές μεταβλητές για τη δημιουργία των συνιστωσών, τον βαθμό δηλαδή στον οποίο είναι, υπεύθυνες γι' αυτές. Η εξίσωση υπολογισμού των φορτίων είναι η κάτωθι: $lij = \frac{a_{ij}}{s_j} * \sqrt{\lambda_i}$ όπου lij είναι το φορτίο της μεταβλητής j για την i συνιστώσα, aij είναι ο συντελεστής στάθμισης της μεταβλητής j για την i συνιστώσα επίσης, λi είναι η χαρακτηριστική ρίζα της i συνιστώσας και sj είναι η τυπική απόκλιση της μεταβλητής j.

Συνοψίζοντας, θα μπορούσαμε να πούμε πως κατά την ανάλυση των κύριων συνιστωσών τα βασικά βήματα που πρέπει να ακολουθηθούν είναι ο μετασχηματισμός των αρχικών μεταβλητών X1, X2,... Xp, έτσι ώστε η διακύμανσή τους να ισούται με τη μονάδα και ο μέσος όρος τους με το απόλυτο 0, ο υπολογισμός της μήτρας των

συσχετίσεων, η αξιολόγηση των συντελεστών στάθμισης aij και των ιδιοτιμών λ1, λ2... λp με σκοπό την απόρριψη των συνιστωσών που συνεισφέρουν μικρό ποσοστό μεταβλητότητας



Εικόνα 7: Ανάλυση Κυρίων Συνιστωσών
Πηγή: https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html/2

2.5.2.2. Μέθοδος Μερικών Ελαχίστων Τετραγώνων (Partial Least Squares)

Η Μέθοδος Μερικών Ελαχίστων Τετραγώνων (PLS) είναι μια στατιστική μέθοδος πολλαπλής γραμμικής παλινδρόμησης που παρουσιάζει πολλά κοινά με την ανάλυση κύριων συνιστωσών (PCA). Ωστόσο έχει φανεί πως ανταποκρίνεται ικανοποιητικότερα σε περιπτώσεις φαινομένων overfitting που η τυπική παλινδρόμηση και η PCA χωλαίνουν. Πρόκειται για περιπτώσεις δηλαδή που η μήτρα των προγνωστικών έχει περισσότερες μεταβλητές από τις παρατηρήσεις οπότε παράγεται ένα μοντέλο που ναι μεν προσαρμόζεται στα πειραματικά δεδομένα, αλλά αποτυγχάνει να προβλέψει τα νέα.

Στόχος της είναι η εύρεση συνιστωσών του X ικανές να προβλέψουν ικανοποιητικά το Y, υπό την προϋπόθεση ότι οι εν λόγω συνιστώσες ερμηνεύουν τη μέγιστη δυνατή συνδιασπορά, καθώς επίσης και η εξαγωγή των λανθάνουσων μεταβλητών (latent values)

οι οποίες ερμηνεύουν την μέγιστη διασπορά στην απόκριση και οδηγούν στην δημιουργία ενός καλού μοντέλου. Οι λανθάνουσες μεταβλητές αποτελούν σημαντικότατο κομμάτι της μεθόδου Μερικών Ελαχίστων Τετραγώνων και γι'αυτό πολλές φορές το ακρωνύμιο PLS μεταφράζεται ως Projection to Latent Structures.[31] Η μερική διακριτή διαίρεση των διαστάσεων των τετραγώνων (PLS-DA) είναι μια παραλλαγή που χρησιμοποιείται όταν το πρόβλημα είναι πρόβλημα κατηγοριοποίησης.

Έστω ότι τα δεδομένα αποτελούνται από έναν πίνακα X που περιλαμβάνει τις ανεξάρτητες μεταβλητές των N παρατηρήσεων και έναν πίνακα Υ που περιλαμβάνει τις εξαρτημένες μεταβλητές των N παρατηρήσεων

Η παλινδρόμηση PLS αποσυνθέτει τόσο το X όσο και το Υ ως προϊόν ενός κοινού συνόλου ορθογωνικών παραγόντων και ένα σύνολο ειδικών φορτίων. Έτσι, οι ανεξάρτητες μεταβλητές αποσυντίθενται ως $X = TP^T$ με $T^TT = I$ με τον I είναι ο πίνακας ταυτότητας (ορισμένες παραλλαγές της τεχνικής δεν απαιτούν το T να έχει πρότυπα μονάδων). Σε αναλογία με την PCA ο T ονομάζεται πίνακας βαθμολογίας, και ο P πίνακας φορτίων (στην PLS τα φορτία δεν είναι ορθογωνικά). Ομοίως, το Υ εκτιμάται ως $\hat{Y} = TBC^T$ όπου το B είναι ένας διαγώνιος πίνακας με τα "βάρη παλινδρόμησης" ως διαγώνια στοιχεία. Οι στήλες του T είναι οι λανθάνουσες μεταβλητές και όταν ο αριθμός τους είναι ίσος με τον βαθμό του X, μπορούμε να πούμε ότι έχει εκτελεστεί μια ακριβής αποσύνθεση του X. Ωστόσο με αυτή την διαδικασία εκτιμάται μόνο το Y, ενώ αξίζει να αναφερθεί ότι γενικά $\hat{Y} \neq Y$. [11]

Όσον αφορά την επιλογή των λανθάνουσων μεταβλητών, μπορεί να πραγματοποιηθεί με πολλούς τρόπους αφού οποιοδήποτε σύνολο ορθογωνικών φορέων που καλύπτουν το χώρο της στήλης του X θα μπορούσε να χρησιμοποιηθεί για να παίξει το ρόλο του T. Για να καθορίσουμε το T, απαιτούνται πρόσθετες συνθήκες. Στην περίπτωση της παλινδρόμησης με την μέθοδο μερικών ελαχίστων τετραγώνων, καθίσταται απαραίτητη η εύρεση δύο συνόλων βαρών w και c προκειμένου να δημιουργηθεί (αντίστοιχα) ένας γραμμικός συνδυασμός των στηλών X και Υ με τρόπο τέτοιο που να μεγιστοποιείται η συνδιακύμανσή τους. Ειδικότερα ο στόχος είναι η απόκτηση ενός

ζεύγους φορέων t = Xw και u = Yc που να συμορφώνεται με τους περιορισμούς ότι $w^Tw = 1$, $t^Tt = 1$ και $t^Tu=max$. Όταν βρεθεί η πρώτη λανθάνουσα μεταβλητή, αφαιρείται από το X και το Υ και η διαδικασία επαναλαμβάνεται έως ότου γίνει ο πίνακας X να γίνει μηδενικός.

Το πρώτο βήμα είναι η δημιουργία δύο πινάκων E = X και F = Y οι οποίοι στη συνέχεια μετασχηματίζονται σε Zscores). Πριν από την έναρξη της διαδικασίας επανάληψης, ο φορέας u αρχικοποιείται με τυχαίες τιμές. (στο εξής όπου χρησιμοποιείται το σύμβολο ∝ σημαίνει "να εξομαλύνει το αποτέλεσμα της λειτουργίας").

Βήμα 1. w ∝ ETu (εκτίμηση των βαρών X).

Βήμα 2. t *α* Ew (εκτίμηση των βαθμολογιών του παράγοντα X).

Βήμα 3. c *α* F Tt (εκτίμηση των βαρών Υ).

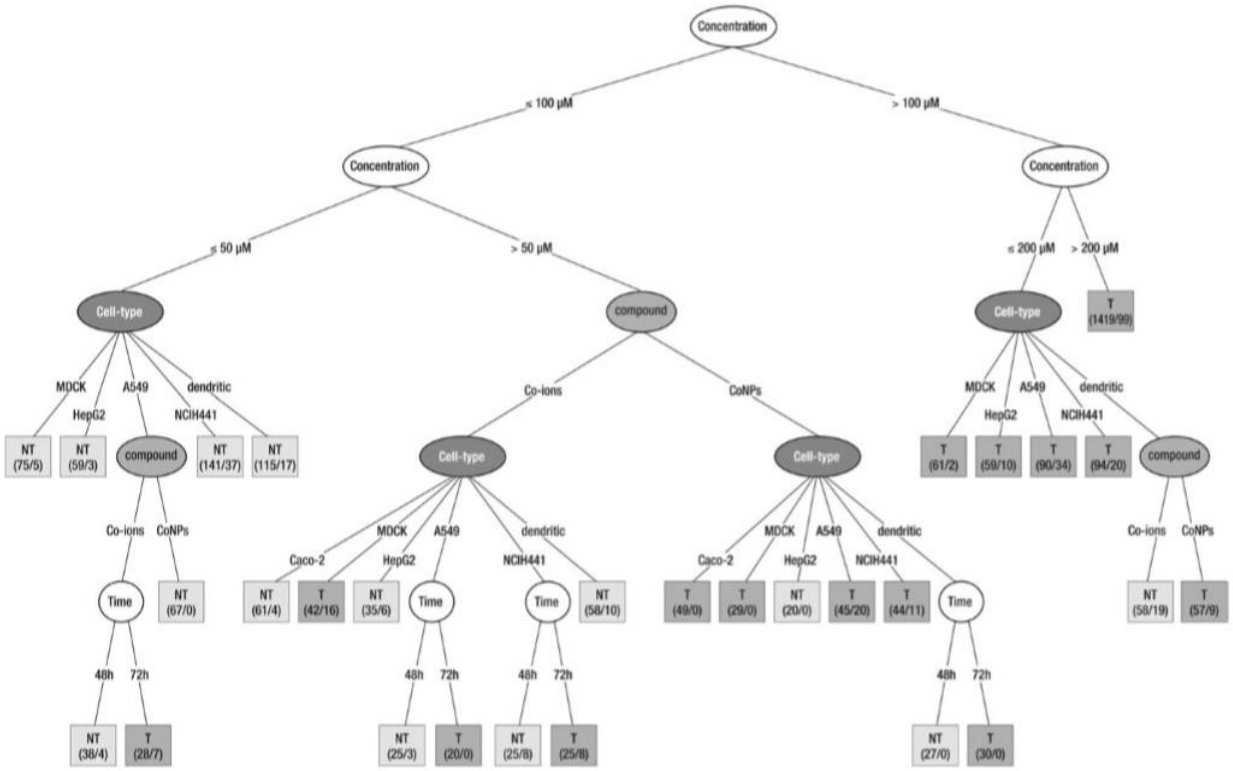
Βήμα 4. u = Fc (εκτίμηση των βαθμολογιών του παράγοντα Υ).

Αν το t δεν έχει συγκλίνει, τότε το βήμα 1επαναλαμβάνεται, εάν το t έχει συγκλίνει, τότε είναι δυνατός ο υπολογισμός της τιμής του b που χρησιμοποιείται για την πρόβλεψη του Y από t ως $b = t^Tu$, και ο υπολογισμός των φορτίων για το X ως $p = E^Tt$. Έπειτα αφαιρείται το αποτέλεσμα του t τόσο από το E όσο και από το F ως εξής $E = E-tp^T$ και $F = F-btc^T$ και οι φορείς t, u, w, c, και p αποθηκεύονται στους αντίστοιχους πίνακες ενώ η βαθμίδα b αποθηκεύεται ως διαγώνιος του B. Το άθροισμα των τετραγώνων του X και αντίστοιχα τουαντίστοιχα Υ εξηγείται από τον λανθάνοντα φορέα υπολογίζεται ως pTp (αντίστοιχα b2), και η εξήγηση διακύμανσης επιτυγχάνεται διαιρώντας το προηγούμενο άθροισμα τετραγώνων με το αντίστοιχο συνολικό άθροισμα τετραγώνων (δηλαδή SSX και SSY). Εάν ο πίνακας E είναι μηδενικός τότε έχει βρεθεί το σύνολο των λανθάνοντων διανυσμάτων, διαφορετικά η διαδικασία επαναλαμβάνεται από το βήμα 1 μέχρι να φτάσουμε στο επιθυμητό αποτέλεσμα.

2.5.3. Δένδρα Αποφάσεων (Decision Trees)

Αποτελούν τον γνωστότερο ίσως αλγόριθμο της επιβλεπόμενης μηχανικής μάθησης και τον πρώτο σε χρήση όταν πρόκειται για προβλήματα ταξινόμησης. Τα δένδρα απόφασης/ταξινόμησης χρησιμοποιούνται για να προβλέψουν,την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών) και οδηγούν στην δημιουργία ενός γραφήματος ροής δενδροειδούς μορφής που με γραφικό τρόπο περιγράφει τα δεομένα, τις εναλλακτικές και τα αποτελέσματα του μοντέλου. Το παραγόμενο σχήμα αποτελείται από κόμβους, κλαδιά και φύλλα. Ο ανώτερος κόμβος του δένδρου ονομάζεται ρίζα και είναι ένας κόμβος χωρίς εισερχόμενα κλαδιά. Κάθε ένας από τους υπόλοιπους κόμβους ορίζει μια συνθήκη ελέγχου της τιμής του εκάστοτε χαρακτηριστικού ενώ κάθε κλαδί που εξέρχεται από κάθε κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή για το χαρακτηριστικό που αντιπροσωπεύει ο κόμβος. Οι κόμβοι που διαθέτουν εξερχόμενα κλαδιά ονομάζονται εσωτερικοί ή κόμβοι εξέτασης ενώ οι υπόλοιποι, δηλαδή τα φύλλα, είναι οι κόμβοι απόφασης ή αλλιώς τερματικοί αφού είναι η απόφαση ταξινόμησης σε κάποια από τις υπάρχουσες κατηγορίες. Συνεπώς τα φύλλα αποτελούν κατηγορίες ομαδοποίησης (classes) και θα μπορούσαμε να πούμε πως ένα δένδρο αποφάσεων είναι επί της ουσίας ένα σύνολο κανόνων ομαδοποίησης (classification rules).

Βασικές προϋποθέσεις για την υλοποίηση ενός αλγορίθμου δένδρου αποφάσεων είναι πρώτα απ' όλα η επιλογή των δεδομένων που θα χρησιμοποιηθούν για την δημιουργία του μοντέλου και έπονται ο καθορισμός, των απαιτούμενων προϋποθέσεων για την δημιουργία του κανόνα ταξινόμησης αλλά και ο καθορισμός των διακριτών κλάσεων ταξινόμησης.



Εικόνα 6: Παράδειγμα Μοντέλου υλοποιημένου με αλγόριθμο Δένδρου Αποφάσεων
Πηγή: Horev-Azaria, L., et al (2011). Predictive toxicology of cobalt nanoparticles and ions: Comparative in vitro study of different cellular models using methods of knowledge discovery from data. Toxicological Sciences, 122(2), 489–501.

2.5.3.1. Ο Αλγόριθμος ID3 (Iterative Dichotomiser 3)

Αποτελεί τον ευρύτατα χρησιμοποιούμενο αλγόριθμο για την δημιουργία δένδρων αποφάσεων. Ο ID3 κατασκευάζει το δένδρο από κάτω προς τα πάνω. Ξεκινά με την επιλογή του καταλληλότερου χαρακτηριστικού για έλεγχο στην ρίζα βάσιζόμενος στις έννοιες της εντροπίας πληροφορίας και κέρδους πληροφορίας και. για κάθε δυνατή τιμή του χαρακτηριστικού δημιουργεί τους αντίστοιχους κόμβους και τα δεδομένα διαμοιράζονται στους σε αυτούς ανάλογα με την τιμή που έχουν για το χαρακτηριστικό που ελέγχεται στη ρίζα. Η όλη διαδικασία επαναλαμβάνεται για κάθε κόμβο που δημιουργείται, δηλαδή, σε κάθε κόμβο του δένδρου αναζητά μεταξύ των χαρακτηριστικών του συνόλου δειγμάτων εκπαίδευσης αυτό το χαρακτηριστικό το οποίο διαχωρίζει

καλύτερα τα δείγματα δεδομένων, εαν βρεθεί χαρακτηριστικό που να διαχωρίζει πλήρως τα δεδομένα εκπαίδευσης, ο αλγόριθμος σταματα, σε αντίθετη περίπτωση συνεχίζει την αναζήτηση στα διαχωρισμένα υποσύνολα ώστε να εντοπίσει το καλύτερο γι' αυτόν χαρακτηριστικό και η διαδικασία ολοκληρώνεται όταν όλοι οι κόμβοι γίνουν τερματικοί, δηλαδή όταν όλα τα δεδομένα που περιέχει έχουν ταξινομηθεί στην ίδια κλάση.

Γίνεται εύκολα κατανοητό ότι η αρχική επιλογή του χαρακτηριστικού για έλεγχο στην ρίζα είναι καθοριστικής σημασίας αφού σε αυτήν στηρίζεται η δημιουργία όλου του δένδρου. Αναφέρθηκε προηγουμένως ότι η επιλογή του κατάλληλου χαρακτηριστικού βασίζεται στις έννοιες της εντροπίας πληροφορίας και κέρδους πληροφορίας. [36]

Η εντροπία πληροφορίας ενός συστήματος δεδομένων σε έναν κόμβο διαχωρισμού ορίζεται με S, εκφράζει την αβεβαιότητα σε ένα σύνολο δεδομένων και υπολογίζεται για δυαδικές περιπτώσεις από την σχέση: $E(S) = -p_+ \cdot \log_2(p_+) - p_- \cdot \log_2(p_-)$ όπου p_+ το κλάσμα των θετικών παραδειγμάτων του συστήματος και p_- το κλάσμα των αρνητικών παραδειγμάτων του συστήματος. Στην μη δυαδική περίπτωση η εντροπία πληροφορίας υπολογίζεται για c αριθμό κατηγοριών από την γενικότερη σχέση: $\mathrm{E}(S) = -\sum_{i=1}^{c} p_i \cdot \log_2(p_i)$

με το p_i να συμβολίζει το ποσοστό των παραδειγμάτων του συνόλου S που ανήκουν στην κατηγορία i. [35]

Η έννοια του κέρδους πληροφορίας A σε ένα σύνολο δεδομένων S, περιγράφει τη συμβολή της μεταβλητής A, εφόσον αυτή επιλεγεί ως χαρακτηριστικό διαχωρισμού, στην μείωση της εντροπίας πληροφορίας του συστήματος. Στόχος είναι η μείωση της εντροπίας πληροφορίας αφού αυτό συνεπάγεται αύξηση στην πυκνότητα πληροφορίας και συνεπώς δένδρο με αυξημένη περιγραφική ικανότητα. Το κέρδος πληροφορίας υπολογίζεται με τη χρήση της εξίσωσης: $G(S,A) = E(S) - \sum_{u \in Values(A)} \frac{|Su|}{|S|} \cdot E(S_u)$

Όπου E(S) : η εντροπία πληροφορίας του προς εξέταση κόμβου

A : η ανεξάρτητη μεταβλητή που λαμβάνει τιμές Values (A)

u : μια από τις τιμές που δύνανται να πάρει το A

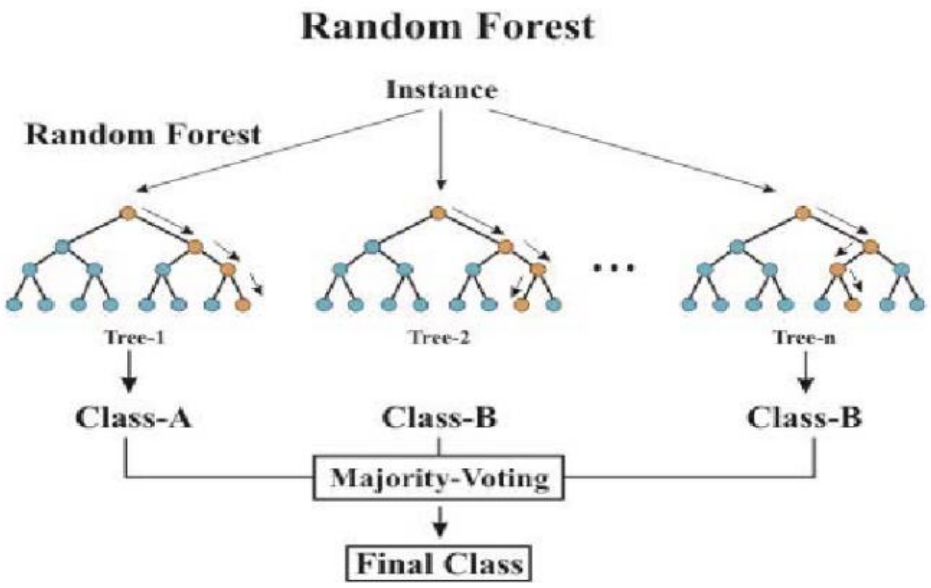
Su : το πλήθος εγγραφών για τις οποίες ισχύει A = u

E(Su) : η εντροπία πληροφορίας του προς εξέταση κόμβου ως προς την τιμή A=u

2.5.4. Τυχαία Δάση (Random Forests)

Τα τυχαία δάση ή αλλιώς τυχαία δάση αποφάσεων είναι μια μέθοδος μηχανικής μάθησης που χρησιμοποιείται τόσο σε προβλήματα παλινδρόμησης όσο και προβλήματα κατηγοριοποίησης/ταξινόμησης και λειτουργεί δημιουργώντας ένα πλήθος δέντρων αποφάσεων με σκοπό την εξαγωγή του αποτελέσματος είτε αυτό είναι η τάξη (ταξινόμηση) είτε αριθμητική πρόβλεψης (παλινδρόμηση) βάση των επιμέρους δένδρων.[12] Τα τυχαία δάση παρομοιάζουν τα δένδρα αποφάσεων, αποδεικνύονται όμως πιο αποτελεσματικά στις περιπτώσεις overfitting όπου τα δένδρα αποφάσεων χολαίνουν.

Ο πρώτος αλγόριθμος για τυχαία δάση αποφάσεων δημιουργήθηκε από τον Tin Kam Ho το 1995 χρησιμοποιώντας την μέθοδο των τυχαίων υποσυνόλων ως ένας τρόπο για την εφαρμογή της προσέγγισης της «στοχαστικής διάκρισης» για την ταξινόμηση που είχε προταθεί από τον Kleinberg λίγα χρόνια νωρίτερα. Στην πάροδο των χρόνων αναπτύχθηκαν βελτιωμένες εκδοχές του αλγορίθμου του Ho με αποκορύφωμα το συνδυασμό της ιδέας του «bagging» με την ιδέα της τυχαίας επιλογής των χαρακτηριστικών που οδήγησε στην δημιουργία μιας συλλογής δέντρων αποφάσεων με ελεγχόμενη διακύμανση και την κατοχύρωση του εμπορικού σήματος «τυχαία δάση» από τους Breiman και Cutler.



Εικόνα 7: Απεικόνιση ταξινομητή υλοποιημένου με αλγόριθμο τυχαίων δασών
Πηγή: **https://www.**semanticscholar**.org**

Γίνεται εύκολα αντιληπτό ότι τα τυχαία δάση αποτελούνται από πολλά δένδρα απόφασης συνεπώς θα έχουν και κάποια κοινά χαρακτηριστικά. Ωστόσο έχουν σημαντικές διαφορές, πρώτα απ' όλα σε αντίθεση με τα κλασσικά δένδρα απόφασης, για την δημιουργία των οποίων αναζητούνται τα χαρακτηριστικά που μεταφέρουν την περισσότερη πληροφορία για να χρησιμοποιηθούν στους πρώτους κόμβους του δένδρου, στα δένδρα που απαρτίζουν ένα τυχαίο δάσος επιλέγεται ένα τυχαίο μικρό υπόδειγμα όλων των χαρακτηριστικών της βάσης, και στην συνέχεια χρησιμοποιούνται τυχαία σε κάθε κόμβο μέχρι να κατασκευαστεί το δένδρο.[39] Μετά την ολοκλήρωση της κατασκευής του, επιλέγεται ένα τμήμα των δεδομένων ως training data για την εκπαίδευσή του και το υπόλοιπο μέρος των δεδομένων αξιοποιείται προς την εκτίμηση του σφάλματος που αποτελεί, όπως σε όλες τις περιπτώσεις μοντελοποίησης, βασικό παράγοντα αξιολόγησης αφού όσο χαμηλότερο είναι το σφάλμα του, τόσο μεγαλύτερη επιρροή ασκεί στο δάσος και συμβάλλει θετικά στο ρυθμό σφάλματος του δάσους. Ένας ακόμη σημαντικός παράγοντας αξιολόγησης της επιτυχίας ενός τυχαίου δάσους είναι η συσχέτιση (ομοιότητα) μεταξύ των δένδρων που το απαρτίζουν. Η ομοιότητα αυτή έγκειται στην

χρήση κοινών χαρακτηριστικών κατά την δημιουργία του και είναι ανάλογη του ρυθμού σφάλματος.

Το Bootstrapping είναι μια τεχνική στατιστικής αναδειγματοληψίας που περιλαμβάνει τυχαία δειγματοληψία ενός συνόλου δεδομένων με αντικατάσταση. Χρησιμοποιείται συχνά ως μέσο ποσοτικοποίησης της αβεβαιότητας που συνδέεται με ένα μοντέλο μηχανικής μάθησης.[8] Η τεχνική του bootstrapping είναι εξαιρετικά χρήσιμη, καθώς επιτρέπει τη δημιουργία νέων δειγμάτων από έναν πληθυσμό χωρίς να χρειάζεται να προχωρήσει και να συλλέξει πρόσθετα «στοιχεία εκπαίδευσης». Η ιδέα πίσω από τη λειτουργία της μεθόδου bootstrap είναι να επαναλαμβάνεται η δειγματοληψία των δεδομένων με αντικατάσταση από το αρχικό σετ κατάρτισης, ώστε να παράγονται πολλαπλά χωριστά σετ εκπαίδευσης. Αυτά χρησιμοποιούνται στη συνέχεια για να επιτρέψουν τις μεθόδους "meta-learner" ή "ensemble" να μειώσουν τη διακύμανση των προβλέψεών τους, βελτιώνοντας έτσι σημαντικά την πρόβλεψη απόδοσης

Εκτός από το Bootstrap, στα δένδρα αποφάσεων και κατά συνέπεια στα τυχαία δάση χρησιμοποιείται το Bootstrap Aggregation ή αλλιώς bagging σκοπός του οποίου είναι η αύξηση της ακρίβειας αλλά και η βελτίωση της σταθερότητας του αλγορίθμου είτε πρόκειται για αλγόριθμο ταξινόμησης είτε για αλγόριθμο παλινδρόμησης. Στα σημαντικά πλεονεκτήματα που παρουσιάζει, συγκαταλέγονται η ικανότητά του να μειώνει σημαντικά τη διακύμανση και η σημαντική συμβολή του στην αποφυγή του φαινομένου overfitting.

Για να αντιληφθούμε τη λειτουργία του αρκεί να θεωρήσουμε ότι έχουμε ένα σύνολο δεδομένων εκπαίδευσης Α μεγέθους κ και από αυτό δημιουργούμε τυχαία σετ εκπαίδευσης A_i επιλέγοντας ομοιόμορφα και με αντικατάσταση από το αρχικό σύνολο Α. Το μέγεθος κάθε συνόλου A_i είναι λ και ισχύει η σχέση κ≤λ. Στην περίπτωση που κ=λ το ποσοστό των στοιχείων που είναι διαφορετικά μεταξύ τους ισούται με το κλάσμα $(1 - \frac{1}{e}) \approx$ 63,2% ενώ όλα τα υπόλοιπα στοιχεία είναι αντίγραφα. Σύμφωνα με τον Breiman, αν και το bagging μπορεί να συμβάλει σημαντικά στην βελτιστοποίηση όταν καλούμαστε να

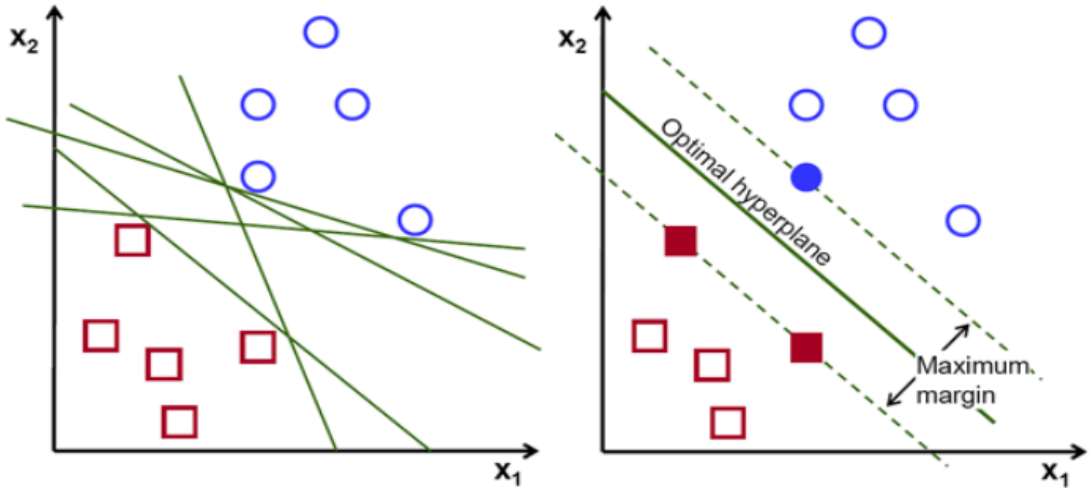
αντιμετωπίσουμε περιπτώσεις ασταθών διαδικασιών, στις περιπτώσεις σταθερών διαδικασιών είναι πιθανόν να προκαλέσει ήπια μεν, υποβάθμιση δε. [3]

2.5.5. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι αλγόριθμοι SVM αν και είναι ίσως η πιο γνωστή κατηγορία αλγορίθμων ταξινόμησης (classification), καθώς παρουσιάζουν αυξημένη ικανότητα γενίκευσης εν συγκρίσει με άλλες παραδοσιακές μεθόδους ταξινόμησης, χρησιμοποιούνται ενίοτε και σε περιπτώσεις προσέγγισης της μορφής της συνάρτησης σε προβλήματα παλινδρόμησης (regression).

Αν επιχειρήσουμε μια σύντομη ιστορική αναδρομή στις Μηχανές Διανυσμάτων Υποστήριξης θα βλέπαμε ότι η συμβολή του Vapnik στην ανάπτυξή τους ήταν μεγάλη. Ξεκίνησε το 1974 οπότε μαζί με τον Chervonenkis διατύπωσαν την θεωρία στατιστικής εκμάθησης την οποία ο Vapnik ανέπτυξε περισσότερο το 1979 διατυπώνοντας πως στόχος της μεθόδου είναι η δημιουργία συνόλων σημείων με ίδιες ιδιότητες (τάξεις/κλάσεις) με υπερπεπίπεδα διαχωρισμού και ακολουθώντας αυτόν τον διαχωρισμό να γίνεται η ταξινόμηση του προς εξέταση στοιχείου στην αντίστοιχη κλάση. Δέκα χρόνια αργότερα, το 1989, οι Anlauf and Biehl που πρότειναν τα υπερεπίπεδα διαχωρισμού με μέγιστο διάκενο, ενώ το 1992 στο συνέδριο COLT οι Boser et al. πρότειναν μια μορφή SVM σχεδόν ίδια με αυτή που γνωρίζουμε και χρησιμοποιούμε σήμερα. Ωστόσο ορόσημο για τα SVM ήταν το 1995 οπότε ο Vapnik υλοποίησε την επέκταση των SVM στα προβλήματα προσέγγισης συνάρτησης (regression).[37]

Οι Μηχανές Διανυσμάτων Υποστήριξης είναι συστήματα που εκπαιδεύονται με τη χρήση αλγορίθμων που βασίζονται στην θεωρία της βελτιστοποίησης και στόχος τους είναι δοθέντος ενός συνόλου δεδομένων με ν-διάστατα διανύσματα η δημιουργία ενός υπερεπιπέδου ως επιφάνεια απόφασης με τρόπο τέτοιο ώστε να μεγιστοποιείται το περιθώριο διαχωρισμού μεταξύ θετικών (+1) και αρνητικών (-1) προτύπων όπως φαίνεται στην παρακάτω εικόνα. [27]



Εικόνα 7: Πιθανά Υπερεπίπεδα
Πηγή: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Σε περιπτώσεις που ο γραμμικός διαχωρισμός των προτύπων δεν είναι εφικτός είναι δυνατή η μεταφορά των δεδομένων σε μεγαλύτερη διάσταση ώστε να είναι δυνατός ο διαχωρισμός τους με βάση το τέχνασμα του πηρύνα (kernel trick). Η μηχανή διανυσμάτων υποστήριξης είναι ένας δυαδικός ταξινομητής, ωστόσο αν απαιτείται διαχωρισμός σε παραπάνω από μια κλάσεις είναι δυνατή χρήση περισσότερων μηχανών διανυσμάτων υποστήριξης και η εφαρμογή διάφορετικών τεχνικών.

Η βασική αρχή που διέπει τις Μηχανές Διανυσμάτων Υποστήριξης και αποτελεί ταυτόχρονα σημαντικό πλεονέκτημά τους έναντι άλλων μεθόδων είναι η ελαχιστοποίηση του κατασκευαστικού ρίσκου (Structural Risk Minimization, SRM) έναντι της ελαχιστοποίησης του εμπειρικού ρίσκου (Empeirical Risk Minimization, ERM) που χρησιμοποιείται από τα νευρωνικά δίκτυα και η οποία εξασφαλίζει καλύτερα αποτελέσματα στην ελαχιστοποίηση του αναμενόμενου ρίσκου. Με βάση την αρχή ελαχιστοποίησης του κατασκευαστικού ρίσκου, η ελαχιστοποίηση του αναμενόμενου ρίσκου προϋποθέτει ταυτόχρονη ελαχιστοποίηση τόσο του εμπειρικού ρίσκου όσο και της VC διάστασης (Vapnik–Chervonenkis (VC) dimension), του ακέραιου δηλαδή αριθμού που οριοθετεί την ικανότητα ενός συστήματος για μάθηση. Αν θεωρήσουμε δηλαδή ένα σύνολο

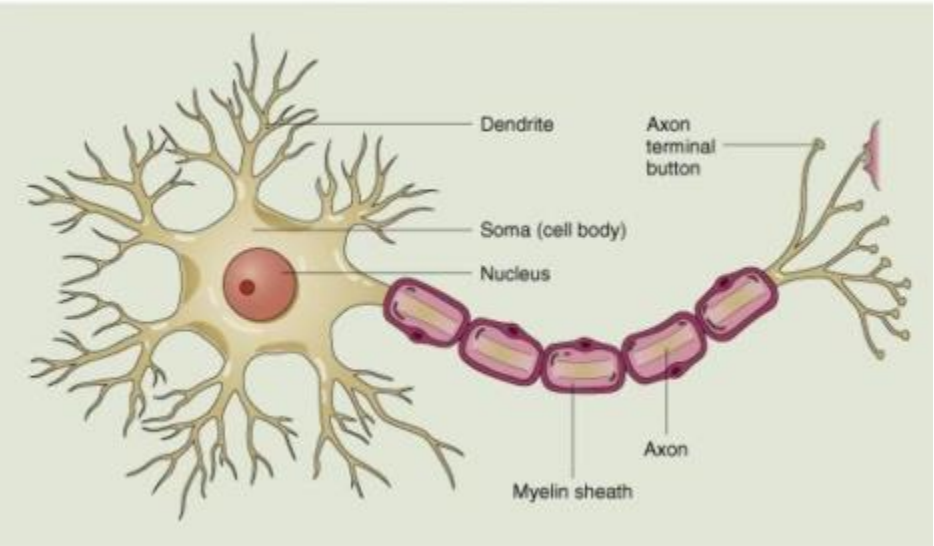
συναρτήσεων {f(α)} τότε ως VC διάσταση ορίζεται μέγιστο πλήθος στοιχείων που μπορούν να αντληθούν από την {f(α)}. Λόγω των σημαντικών δυνατοτήτων τους και της σημαντικής ακρίβειας που προσφέρουν στην πρόβλεψη, οι αλγόριθμοι SVM χρησιμοποιούνται σε διάφορους τομείς της επιστήμης που καλούνται να επεξεργαστούν δεδομένα με κυριότερο αυτόν της βιολογίας - βιοχημείας και της ιατρικής (π.χ. ταξινόμηση πρωτεϊνών).

2.5.6. Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

Τα νευρωνικό δίκτυο είναι αλγόριθμος που χρησιμοποιείται κατά κόρον στη μηχανική μάθηση και η δημιουργία του ξεκίνησε από την υπόθεση ότι ο ανθρώπινος εγκέφαλος είναι ένας μικρός μεν, σύνθετος δε, ηλεκτρονικός υπολογιστής που είναι σε θέση να αξιοποιεί τα δεδομένα εισόδου που δέχεται για την δημιουργία δεδομένων εξόδου (εντολών). Τα εξελιγμένα νευρωνικά δίκτυα είναι εργαλεία που χρησιμοποιούνται για την μη γραμμική μοντελοποίηση σύνθετων σχέσεων μεταξύ δεδομένων εισόδου και εξόδου μέσω μιας ομάδας διασυνδεδεμένων τεχνητών νευρώνων που επεξεργάζονται τις εισερχόμενες πληροφορίες και επικοινωνώντας μεταξύ τους εκτελούν τους επιθυμητούς υπολογισμούς. Παρά το γεγονός ότι τα νευρωνικά δίκτυα ξεκίνησαν ως μια προσπάθεια προσομοίωσης του ανθρώπινου εγκεφάλου και της εξέλιξής του, η εξέλιξή τους σήμερα είναι σχεδόν ανεξάρτητη παρότι εξακολουθεί να βασίζεται στις σημαντικές ομοιότητες που παρατηρούνται μεταξύ βιολογικών και τεχνητών νευρώνων.[45]

Αναφέρθηκε ήδη πως η λειτουργία των νευρωνικών δικτύων βασίζεται στην επικοινωνία των τεχνητών νευρώνων, η δημιουργία των οποίων προσομοιάζει τους βιολογικούς νευρώνες, τα κύρια δομικά στοιχεία του εγκεφάλου, που αποτελούνται από το σώμα, τον άξονα και τους δενδρίτες Η βασική λειτουργία του βιολογικού νευρώνα είναι η μεταφορά σημάτων στους υπόλοιπους νευρώνες μέσω του νευρωνικού δικτύου. Το βιολογικό νευρωνικό δίκτυο αποτελείται από πολλούς διασυνδεδεμένους, μέσω των δενδριτών, νευρώνες οι οποίοι αφού δεχθούν και επεξεργαστούν τα ερεθίσματα που

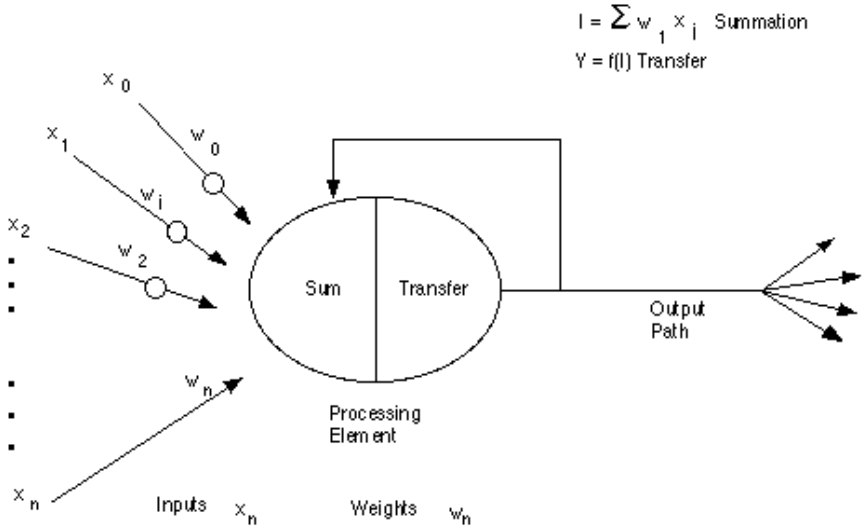
δέχονται τα αποτελούν το αποτέλεσμα της επεξεργασίας μέσω των συνάψεων, δηλαδή του σημείου των χημικών διεργασιών, όπου ο άξονας ενός νευρώνα μεταδίδει το σήμα στους δενδρίτες των επόμενων νευρώνων και έτσι ο άνθρωπος είναι σε θέση να επιτελέσει λειτουργίες ανάλογα με τα προσλαμβανόμενα ερεθίσματα.



Εικόνα 8: Βιολογικός Νευρώνας
Πηγή: www.google.com

Αντίστοιχα με τον βιολογικό νευρώνα, ένας τεχνητός νευρώνας δέχεται τα ερεθίσματα για το καθένα από το οποίο έχει οριστεί κάποια βαρύτητα και αφού υπολογίσει το άθροισμα των βεβαρυμένων ερεθισμάτων μεταφέρει στους επόμενους τεχνητούς νευρώνες το ερέθισμα που προκύπτει από τη συνάρτηση μεταφοράς του νευρώνα. Η λειτουργία του μοντέλου του τεχνητού νευρώνα γίνεται πιο ξεκάθαρη με μια ματιά στην παρακάτω εικόνα. Τα x0, x1, … αποτελούν τα ερεθίσματα που λαμβάνονται από τους προηγούμενους νευρώνες και τα σημεία w0, w1, … απεικονίζουν τους αντίστοιχους συντελεστές βαρύτητας των ερεθισμάτων. Η άθροιση των των βεβαρημένων ερεθισμάτων λαμβάνει χώρα στον πυρήνα σύμφωνα με την εξίσωση: $F = \sum_{i}^{n} x_i w_i$ και βασιζόμενος στην συνάρτηση μεταφοράς-μετάβασης, όταν το ζυγισμένο άθροισμα των εισόδων είναι μεγαλύτερο από την τιμή ενεργοποίησης (threshold value) ο νευρώνας

δημιουργεί το ερέθισμα εξόδου που προωθείται στους επόμενους νευρώνες, δηλαδή όταν $\sum_{i}^{n} x_i w_i - \theta > 0$ όπου θ: threshold value.



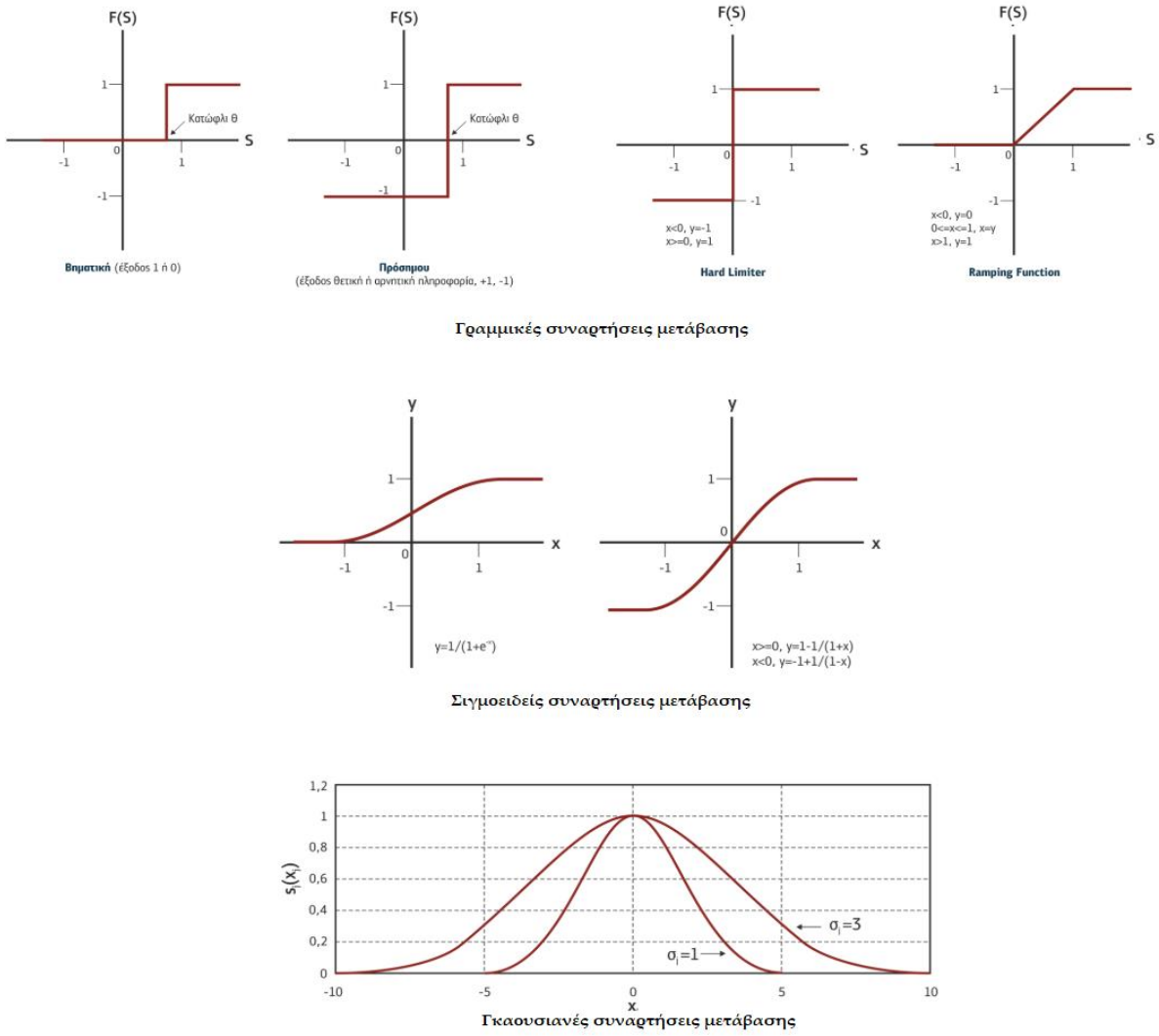
Εικόνα 9: Αναπαράσταση Τεχνητού Νευρώνα
Πηγή: https://nucleus2012.wordpress.com

Το απλούστερο νευρωνικό δίκτυο είναι ο νευρώνας Perceptron που αποτελείται από έναν μόνο νευρώνα και εφευρέθηκε το 1957 στο Αεροναυτικό Εργαστήριο του Κορνέλλ από τον Φρανκ Ρόζενμπλαττ (Frank Rosenblatt). Στην περίπτωση λοιπόν ενός διανύσματος εισόδου $x=(x_1, x_2,.., x_n)$ η έξοδος του Perceptron υπολογίζεται από την δυαδική γραμμική συνάρτηση μετάβασης g με βάση την εξίσωση:

$$a = g(\sum_{i}^{n} x_i w_i)$$

Αν και οι απλούστερες συναρτήσεις μετάβασης είναι οι γραμμικές, όπως π.χ. οι βηματικές ή συναρτήσεις κατωφλίου (threshold functions), οι συναρτήσεις προσήμου (sign functions), οι συναρτήσεις βηματικής μεταβολής (hard limiter functions), οι συναρτήσεις αναρρίχησης (ramping functions) ή η δυαδική γραμμική που χρησιμοποιεί ο νευρώνας Perceptron, είναι γεγονός ότι στις περισσότερες περιπτώσεις δεν επιλέγονται αυτές αλλά οι

μη γραμμικές συναρτήσεις, όπως οι σιγμοειδείς (sigmoid functions) και οι Γκαουσιανές συναρτήσεις (Gaussian functions) οι οποίες είναι αρκετά πιο πολύπλοκες αλλά εξασφαλίζουν καλύτερα αποτελέσματα.

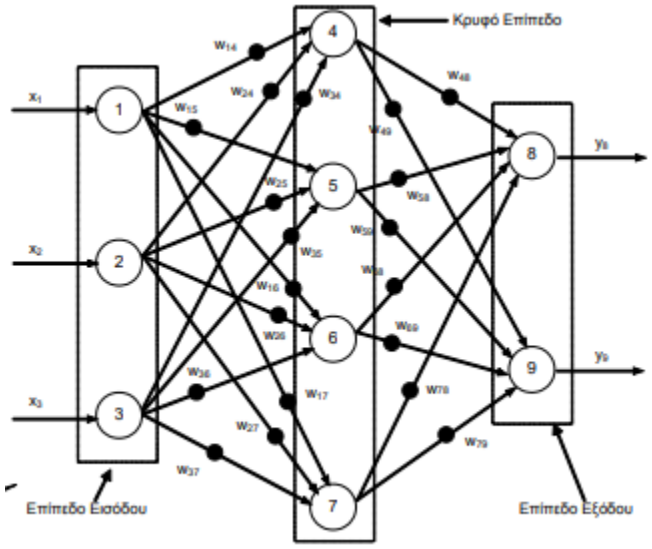


Εικόνα 10: Γραφική αναπαράσταση των συχνότερα χρησιμοποιούμενων συναρτήσεων μεταφοράς
Πηγή: http://repfiles.kallipos.gr/html_books/93/index.html

Ένα νευρωνικό δίκτυο προκύπτει από τον συνδυασμό πολλών τεχνητών νευρώνων διασυνδεδεμένων με συναπτικές συνδέσεις και συναρτήσεις ενεργοποίησης και

αποτελείται από τρία επίπεδα, το επίπεδο εισόδου (input layer), το επίπεδο εξόδου (output layer) και το κρυμμένο επίπεδο (hidden layer) που δεν είναι ορατό από τα άλλα επίπεδα. Τα επίπεδα αυτά περιέχουν αντίστοιχα τους νευρώνες εισόδου, τους νευρώνες εξόδου και τους υπολογιστικούς ή κρυμμένους νευρώνες και τα σήματα εξόδου του ενός επιπέδου χρησιμοποιούνται για την ανατροφοδότηση του επόμενου επιπέδου. Ενώ αξίζει να αναφερθεί ότι τα νευρωνικά δίκτυα δεν διαθέτουν απαραίτητα μόνο ένα κρυφό επίπεδο αφού είναι δυνατή η ύπαρξη και παραπάνω.

Ο διαχωρισμός των νευρωνικών δικτύων πραγματοποιείται τόσο με βάση τον τρόπο σύνδεσης των νευρώνων όσο με βάση την ροή της πληροφορίας. Όσον αφορά την πρώτη κατηγοριοποίηση τα νευρωνικά δίκτυα χωρίζονται σε πλήρως συνδεδεμένα (fully connected) στις περιπτώσεις που όλοι οι νευρώνες του ενός επιπέδου συνδέονται με όλους τους νευρώνες του επόμενου επιπέδου και σε μερικώς συνδεδεμένα στις υπόλοιπες περιπτώσεις.



Εικόνα 10: Τεχνητό Νευρωνικό Δίκτυο απλής τροφοδότησης 3-4-2 Πλήρως συνδεδεμένο
Πηγή: Ιώαννης Βλαχάβας, Τεχνητή Νοημοσύνη, Τεχνητά Νευρωνικά Δίκτυα, Τμήμα Πληροφορικής ΑΠΘ, Θεσσαλονίκη 2013

Ο επόμενος διαχωρισμός έγκειται στην ροή πληροφορίας και έτσι έχουμε τα δίκτυα πρόσθιας τροφοδότησης (feedforward) και τα δίκτυα με ανατροφοδότηση (feedback ή recurrent). Στην πρώτη κατηγορία (feedforward) η ροή της πληροφορίας είναι μιας κατεύθυνσης, πάντα προς τα εμπρός και δεν είναι δυνατή η προς τα πίσω τροφοδότηση του δικτύου αφού δεν υπάρχουν νευρώνες σύνδεσης από τους νευρώνες ενός επιπέδου προς νευρώνες προηγούμενου επιπέδου ενώ στην δεύερη (feedback ή recurrent) οι νευρώνες ενός επιπέδου όχι μόνο συνδέονται με νευρώνες προηγούμενου επιπέδου αλλά είναι δυνατόν να συνδέονται και με άλλους νευρώνες του ίδιου επιπέδου. Στα νευρωνικά δίκτυα με ανατροφοδότηση οι υπολογισμοί πραγματοποιούνται σε δύο φάσεις, στην πρώτη πραγματοποιούνται οι υπολογισμοί που αφορούν την πρόσθια τροφοδότηση και στην δεύτερη οι υπολογισμοί που αφορούν τις συνδέσεις ανατροφοδότησης. Αν και η μεγαλύτερη πλειοψηφία των εφαρμογών νευρωνικών δικτύων χρησιμοποιούνται δίκτυα πρόσθιας τροφοδότησης διότι σε αυτά οι αλγόριθμοι μάθησης είναι πιο ισχυροί, υπάρχουν αρκετά προβλήματα που αυτά δεν μπορούν να αντιμετωπίσουν και έτσι η χρήση δικτύων με ανατροφοδότηση καθίσταται μονόδρομος μιας που αυτά είναι πολύ πιο κοντά στα βιολογικά νευρωνικά δίκτυα από τα feedforward. [45]

Αφού λοιπον καθοριστούν η δομή του δικτύου, δηλαδή ο τρόπος σύνδεσης των νευρώνων και ο τύπος ροής της πληροφορίας, αν δηλαδή είναι πρόσθιας τροφοδότησης ή με ανατροφοδότηση μένουν η εκπαίδευση και η ανάκληση του ώστε να ολοκληρωθεί. Στο στάδιο της εκπαίδευσης πραγματοποιείται τροποποίηση της τιμ''ης των βαρών του δικτύου ώστε για συγκεκριμένο διάνυσμα εισόδου να παράγεται ως αποτέλεσμα συγκεκριμένο διάνυσμα εξόδου.[35] Στην συνέχεια ακολουθεί η φάση της ανάκλησης η οποία πραγματοποιείται με τη χρήση ενός νέου συνόλου δειγμάτων τα οποία δεν είχαν πάρει μέρος στη διαδικασία εκπαίδευσης. Η τροποποίηση των βαρών ένος νευρωνικού δικτύου γίνεται είτε μέσω επιβλεπόμενης μάθησης είτε μέσω μη επιβλεπόμενης. Στην πρώτη περίπτωση το δίκτυο τροφοδοτείται τόσο με τιμές εισόδου όσο και με τιμές εξόδου (επιθυμητές) και βασιζόμενο στην τρέχουσα κατάσταση βαρών παράγει μια έξοδο η οποία παρουσιάζει απόκλιση από την επιθυμητή (error). Η αναπροσαρμογή των βαρών γίνεται κυρίως αξιοποιώντας το error και με τη χρήση κατάλληλου αλγόριθμου μάθησης όπως

είναι ο Κανόνας Δέλτα (Delta rule learning), ο αλγόριθμος αντίστροφης μετάδοσης λάθους (back propagation), η ανταγωνιστική μάθηση (competitive learning) ή η τυχαία μάθηση (random learning. Στην δεύτερη περίπτωση επικρατεί η ικανότητα του συστήματος να αναδιοργανώνεται από μόνο του με βάσει τα δεδομένα εισόδου που λαμβάνει παράδειγμα αποτελούν τα δίκτυα Kohonen, χαρακτηριστικό των οποίων είναι η δυνατότητα να ταξινομούν διανύσματα με την βοήθεια ενός αλγόριθμου μη επιβλεπόμενης μάθησης και να οργανώνουν τον πίνακα των βαρών τους (w) με τέτοιο τρόπο ώστε αναγνωρίζει όποια κανονικότητα μπορεί να υπάρχει στα διανύσματα εισόδου. [50]

2.5.6.1. Αυτοοργανούμενος Χάρτης (self-organizing map,SOM)

Ο αυτοοργανούμενος χάρτης (SOM) είναι ένας τύπος τεχνητού νευρικού δικτύου (ANN) που εκπαιδεύεται χρησιμοποιώντας μη επιβλεπόμενη μάθηση για να παράγει μια χαμηλών διαστάσεων (συνήθως δισδιάστατη) αναπαράσταση των προτύπων εισόδου των δειγμάτων κατάρτισης, που ονομάζεται χάρτης, και μπορεί να χρησιμοποιηθεί στην λήψη αποφάσεων. Παρά το γεγονός ότι παρουσιάζει σημαντικές ομοιότητες με τα δίκτυα εμπρόσθιας τροφοδότησης feedforward, ο αυτοοργανούμενος χάρτης διαφέρει τόσο σε διάταξη όσο σε κίνητρο. Επίσης διαφέρει καθώς εφαρμόζει ανταγωνιστική μάθηση, σε αντίθεση με τη μάθηση διόρθωσης σφαλμάτων που εφαρμόζεται από τα άλλα τεχνητά νευρωνικά δίκτυα, εννοώντας ότι οι νευρώνες εξόδου ανταγωνίζονται μεταξύ τους για το δικαίωμα ενεργοποίησης με αποτέλεσμα μόνο ένας νευρώνας να καθίσταται ενεργός κάθε στιγμή. Ο δημοφιλέστερος αυτοοργανούμενος χάρτης είναι το τεχνητό νευρωνικό δίκτυο που εισήγαγε ο φινλανδός καθηγητής Teuvo Kohonen το 1982.

Όπως και τα περισσότερα τεχνητά νευρωνικά δίκτυα, τα SOM λειτουργούν σε δύο φάσεις: το training και το mapping. Το training δημιουργεί τον χάρτη αξιοποιώντας τα δεδομένα εισόδου, ενώ το mapping ταξινομεί αυτόματα ένα νέο διάνυσμα εισόδου. Το ορατό τμήμα ενός αυτο-οργανωτικού χάρτη ε, που αποτελείται από νευρώνες. Σύμφωνα με την άποψη που διατύπωσε το 1996 ο Jaakko Hollmen, ο χάρτης είναι συνήθως μια

δισδιάστατη περιοχή όπου οι κόμβοι είναι διατεταγμένοι σε κανονικό εξαγωνικό ή ορθογώνιο πλέγμα και κάθε κόμβος συνδέεται με ένα διάνυσμα "βάρους", το οποίο υποδηλώνει μια θέση στον χώρο εισόδου. Στόχος είναι η μείωση της απόστασης μεταξυ των διανυσμάτων των βαρών και των δεδομένων εισόδου, με παράλληλη διατήρηση της τοπολογίας που δημιουργεί ο χάρτης. Θα μπορούσαμε λοιπόν να πούμε ότι ένας αυτοογανούμενος χάρτης αντιστοιχίζει έναν χώρο υψηλών διαστάσεων σε έναν χάρτη χαμηλότερων διαστάσεων ικανό να ταξινομει ένα εισερχόμενο διάνυσμα στον κόμβο με το πλησιέστερο διάνυσμα βάρους προς αυτό.

Συνοψίζοντας θα μπορούσαμε να πούμε πως τα νευρωνικά δίκτυα είναι ικανά να μοντελοποιήσουν εξαιρετικά πολύπλοκες λειτουργίες στις οποίες αποτυγχάνουν οι παραδοσιακές μέθοδοι. Παρουσιάζουν σημαντική ανοχή σε δεδομένα εκπαίδευσης με θόρυβο, δηλαδή δεδομένα ενδεχομένως περιέχουνλανθασμένες τιμές, λόγου χάρη από λανθασμένη καταχώρηση, ενώ δεν απαιτούν εξειδικευμένο επίπεδο γνώσεων από τον χρήστη αφού αρκεί να συγκεντρώσει τα δεδομένα και στη συνέχεια να τα τροφοδοτήσει στο κατάλληλο νευρωνικό δίκτυο το οποίο «αντιλαμβάνεται» αυτομάτως τη δομή των δεδομένων και την μεταφράζει σε κατάλληλες επιλογές συναπτικών βαρών. Σημαντικό μειονέκτημα ωστόσο αποτελεί η αδυναμία ποιοτικής ερμηνείας της μοντελοποιούμενης γνώσης. Παρόλα αυτά όμως τα νευρωνικά δίκτυα χρησιμοποιούνται σε όλο σχεδόν το φάσμα της επιστήμης, ιατρική, μαθηματικά, οικονομικά, μηχανική λόγω των σπουδαίων ιδιοτήτων τους.

2.5.7. Ταξινομητές Naive-Bayes (Naive-Bayes classifiers)

Οι Μπεϋζιανοί ταξινομητές χρησιμοποιούνται για την μοντελοποίηση μιας πιθανοτικής σχέσης μεταξύ των χαρακτηριστικών και της κατηγορίας,[49] στόχος τους είναι δηλαδή η πρόβλεψη της πιθανότητας ένα δείγμα Α να ανήκει σε κάποια κατηγορία (π.χ Β). Ο απλούστερος Bayesian κατηγοριοποιητής είναι ο Naïve Bayesian. Αυτός υποθέτει ότι η επίδραση ενός γνωρίσματος σε μία κατηγορία είναι ανεξάρτητη από τις τιμές των

υπόλοιπων γνωρισμάτων. Ο λόγος που γίνεται αυτό είναι για να αποφεύγονται οι πολύπλοκοι υπολογισμοί κατά τη συνθήκη ανεξαρτησίας της κατηγορίας. Η αρχή στην οποία βασίζεται η λειτουργία των Μπεϋζιανών ταξινομητών είναι το θεώρημα του Bayes το οποίο περιγράφεται από την εξίσωση:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

όπου: P(A|B): η πιθανότητα να πραγματοποιηθεί το Α με δεδομένο ότι συμβαίνει το Β, αυτή ονομάζεται εκ των υστέρον πιθανότητα (posterior probability)

P(B|A): η πιθανότητα του δεδομένου Β με δεδομένο ότι η υπόθεση Α ήταν αληθής

P(A): η πιθανότητα η υπόθεση Α να είναι αληθής ανεξάρτητα από τα δεδομένα, αυτή ονομάζεται εκ των προτέρον πιθανότητα (posterior probability)

P(B): η πιθανότητα του δεδομένου Β ανεξάρτητα από την υπόθεση Α

Συνεπώς προκύπτει πως αν είναι γνωστές οι τιμές των χαρακτηριστικών ενός παραδείγματος, ο Μπεϋζιανός ταξινομητής είναι σε θέση να υπολογίσει τις υπό συνθήκη πιθανότητες ώστε να ανήκει σε όλες τις πιθανές κατηγορίες. Επειδή όμως είναι εξαιρετικά δύσκολο για τον ταξινομητή να υπολογίσει τις υπό συνθήκη πιθανότητες εκτός και αν τα δεδομένα εκπαίδευσης του αλγόριθμου καλύπτουν πλήρως τα χαρακτηριστικά, κάτι που όμως δεν συμβαίνει συχνά, κρίνεται απαραίτητο να υιοθετηθούν άλλες πρακτικές και προσεγγίσεις. Στα πλαίσια αυτά η γενική αρχή που διέπει την παραπάνω εξίσωση είναι αυτή της ανεξαρτησίας των χαρακτηριστικών και η παραδοχή ότι η παρουσία του ενός δεν επηρεάζει το άλλο.

2.5.7.1. Gaussian Naive Bayes

Οι ταξινομητές Naive Bayes μπορούν να επεκταθούν και να παρέχουν πραγματικά αξιόπιστα αποτελέσματα, αν υποθέσουμε ότι τα δεδομένα εκπαίδευσης ακολουθούν

κατανομή Gauss. Αυτή η επέκταση ονομάζεται **Gaussian Naive Bayes** και είναι η πιο εύκολη από τις υπόλοιπες προσεγγίσεις καθώς οι μόνοι υπολογισμοί που απαιτεί είναι αυτοί της μέσης τιμής και της τυπικής απόκλισης από τα δεδομένα εκπαίδευσης. Για παράδειγμα, ας υποθέσουμε ότι τα δεδομένα εκπαίδευσης περιέχουν ένα συνεχές χαρακτηριστικό, x. Κατατάσσουμε πρώτα τα δεδομένα από την κλάση και στη συνέχεια υπολογίζουμε τη μέση τιμή και τη διακύμανση του x σε κάθε κλάση. Έστω ότι για την κλάση C_k ο μέσος όρος των τιμών στο x που σχετίζονται με την κλάση Ck είναι μ_k και η διακύμανση των τιμών στο στο x που σχετίζονται με την κλάση Ck είναι $\sigma^2 k$. Ας υποθέσουμε ότι έχουμε συγκεντρώσει κάποια τιμή παρατήρησης v. Η κατανομή πιθανότητας v που δίνεται σε μια τάξη C_k μπορεί να υπολογιστεί με την προσάρτηση του v στην εξίσωση για μια κανονική κατανομή παραμετροποιημένη από την μέση τιμή και την απόκλιση όπως φαίνεται παρακάτω:

$$p = (x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

2.5.7.2. Multinomial Naive Bayes

Μια άλλη κατηγορία μπεϋζιανών ταξινομητών είναι οι multinomial ταξινομητές κατά Bayes. Αυτοί διαφέρουν σημαντικά από τους Gaussian αφού δεν είναι δυνατή η εφαρμογή τους σε περιπτώσεις συνεχών δεδομένων αλλά χρησιμοποιούνται πολύ συχνά σε προβλήματα που σκοπό έχουν την ταξινόμηση κειμένου.

Μέσω ενός πολυωνυμικού μοντέλου, τα δείγματα αντιπροσωπεύουν τις συχνότητες με τις οποίες έχουν δημιουργηθεί ορισμένα γεγονότα από ένα πολυώνυμο (p_1, p_2, ….., p_n), όπου p_i είναι η πιθανότητα εμφάνισης του συμβάντος i. Τα διανύσματα **x** = ($x1$, $x2$, … , xn) είναι ιστογράμματα όπου κάθε χαρακτηριστικό εκφράζει συχνότητα εμφάνισης ενός ενδεχομένου i. Η πιθανότητα παρατήρησης ιστογράμματος x δίνεται από την εξίσωση:

$$p(x \mid C_k) = \frac{(\Sigma iCi)!}{\Pi ix!} \Pi {p_{ki}}^{xi}$$

2.5.7.3. Bernoulli Naive Bayes

Μια ακόμα παραλλαγή των Μπεϋζιανών ταξινομητών είναι το μοντέλο Bernoulli Naive Bayes, που όπως και το προαναφερόμενο multinomial χρησιμοποιείται ευρέως σε εργασίες ταξινόμησης κειμένων. Η διαφορά τους έγκειται στο ότι εδώ οι μεταβλητές που χρησιμοποιούνται είναι δυαδικές, εστιάζουν δηλαδή στην εμφάνιση ή όχι ενός χαρακτηριστικού, αδιαφορώντας για την συχνότητα εμφάνισης του. Αν λοιπόν ορίσουμε ως x την δυαδική μεταβλητή που καθορίζει την ύπαρξη ή την μη ύπαρξη ενός χαρακτηριστικού i, τότε η πιθανότητα να ανήκει σε μια κλάση C_k δίνεται από την εξίσωση:

$$p(x \mid C_k) = \prod_{i=1}^{n} p_{ki}^{xi} (1 - p_{k_i})^{(1-x_i)}$$

όπου p_{k_i} : η πιθανότητα η κλάση C_k να περιέχει το χαρακτηριστικό x.

Ο συγκεκριμένος ταξινομητής είναι ιδανικός για ταξινόμηση μικρής έκτασης εγγράφων αφού λόγω των δυαδικών μεταβλητών έχει το πλεονέκτημα μοντελοποίησης παρουσίας ή απουσίας όρων.

3. Τεχνικές QSAR

3.1. Γενικά

Οι κυριότερες μέθοδοι που χρησιμοποιούνται εδώ και χρόνια από την επιστημονική κοινότητα των βιολογικών επιστημών χωρίζονται σε δύο κατηγορίες, τις in vivo μελέτες και τις in vitro μελέτες. Τα τελευταία χρόνια όμως μετά τις ραγδαίες εξελίξεις στον τομέα της επιστήμης των υπολογιστών έχει κερδίσει ιδιαίτερο έδαφος μια νέα σχετικά κατηγορία μελετών, οι in silico μέθοδοι. Ακολουθεί μια σύντομη επισκόπηση των προαναφερόμενων κατηγοριών που σκοπό έχει την ανάδειξη των πλεονεκτημάτων των in silico μεθόδων.

3.1.1. In vivo

Πρόκειται για λατινική έκφραση που στα ελληνικά αποδίδεται «εν ζωή», αναφέρεται λοιπόν σε μελέτες και δοκιμές που πραγματοποιούνται σε κύτταρα, ιστούς ή όργανα έμβιων οργανισμών. Με την καθιέρωση των εναλλακτικών μεθόδων, όρος που επινοήθηκε το 1978 από τον David Smyth και αφορά εκείνες τις μεθόδους που είναι σε θέση να δώσουν τις ίδιες πληροφορίες με τις συμβατικές, χωρίς να χρησιμοποιούν ζώα, μειώνοντας τον αριθμό τους, ή βελτιώνοντας τις συνθήκες πειραματισμού όταν αναφερόμαστε σε in vivo τεχνικές αναφερόμαστε κυρίως σε τεχνικές που περιλαμβάνουν φυλογενετικά κατώτερους οργανισμούς όπως έντομα, μαλάκια ή αμφίβια.[46] Οι τεχνικές αυτές παρά το γεγονός ότι έχουν κάποια πλεονεκτήματα, παρουσιάζουν και αρκετά μεγάλη απόσταση από το «σύστημα ενδιαφέροντος» που είναι ο άνθρωπος [46] και αντίκειται και στους κανόνες της βιοηθικής.

3.1.2. In vitro

Λατινική έκφραση που στα ελληνικά σημαίνει «στο γυαλί» κοινώς στον δοκιμαστικό σωλήνα και αναφέρεται σε τεχνικές και μεθόδους που πραγματοποιούνται έξω από τους ζωντανούς οργανισμούς και περιλαμβάνουν συνήθως καλλιέργειες

κυττάρων, ιστών ή οργάνων. Σε αντίθεση με τις in vivo τεχνικές δίνουν τη δυνατότητα πραγματοποίησης του πειράματος σε αυστηρά ελεγχόμενες συνθήκες και παρουσιάζουν αρκετά πλεονεκτήματα, πέρα από το προφανές που είναι η μη χρήση έμβιων οργανισμών. Στα πλεονεκτήματα αυτά συγκαταλέγονται η δυνατότητα να μελετηθεί σε βάθος το υπό εξέταση φαινόμενο, η δυνατότητα πραγματοποίησης μεγάλου αριθμού επαναλήψεων, το μικρό κόστος και η ήδη αναφερθείσα δυνατότητα για πραγματοποίηση του πειράματος σε αυστηρά ελεγχόμενες συνθήκες. Άριστος έλεγχος των συνθηκών της υπό μελέτη διαδικασίας. Παρόλα αυτά η απουσία του οργανισμού έχει ως αποτέλεσμα την απουσία διασυστημικών και ενδοσυστημικών αλληλεπιδράσεων που θα οδηγούσε στην εξαγωγή πληρέστερου συμπεράσματος αποτελεί σημαντικό μειονέκτημα αυτών των τεχνικών [46]

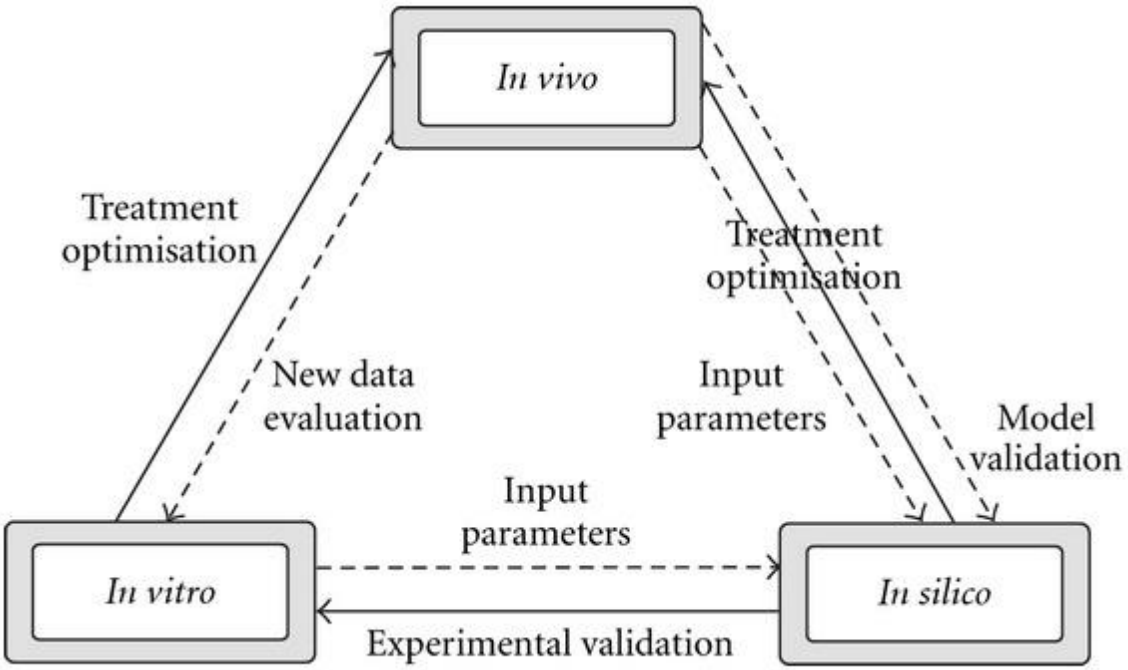
3.1.3. In silico

Η έκφραση in silico χρησιμοποιήθηκε για πρώτη φορά το 1989 στο συμπόσιο «Cellular Automata: Theory and Applications» [54] στο Λος Άλαμος του Νέου Μεξικού όπου ένας μαθηματικός, ονόματι Pedro Miramontes, από το Εθνικό Ανεξάρτητο Πανεπιστήμιο του Νέου Μεξικού κατά την παρουσιάση μιας αναφοράς με τίτλο: *«DNA και RNA: Φυσικοχημικοί περιορισμοί, κυτταρικά αυτόματα και μοριακή εξέλιξη»* χρησιμοποίησε την φράση in silico για να περιγράψει βιολογικά πειράματα που πραγματοποιήθηκαν με τεχνικές μέσω υπολογιστή ή μέσω προσομοίωσης σε υπολογιστή.[19] Οι τεχνικές in silico περιλαμβάνουν αφενός μεθόδους που βασίζονται στην ανάλυση σχέσης δομής-δράσης μορίων και αφετέρου μοντέλα προσομοίωσης δράσεων, παρεμβάσεων και εξέλιξης. [46]

Η χρήση μαθηματικών μοντέλων προσομοίωσης στον υπολογιστή χρησιμοποιείται από πολλούς τομείς της επιστήμης και κατά κόρων από την ιατρική και την φαρμακοβιομηχανία για την ανάπτυξη νέων μεθόδων στη διάγνωση και την αντιμετώπιση ασθενειών και των σχεδιασμό νέων καινοτόμων φαρμάκων αντίστοιχα λόγω των σημαντικών πλεονεκτήματων που παρουσιάζει. Στα πλεονεκτήματα αυτά περιλαμβάνονται: [16]

- Το γεγονός ότι οι παράμετροι εισόδου μπορούν εύκολα να αλλάξουν και να δώσουν γρήγορα νέα αποτελέσματα.
- Το ότι διάφοροι μηχανισμοί μπορούν να μελετηθούν μεμονωμένα, προσδιορίζοντας τον αντίκτυπό τους σε συγκεκριμένες διαδικασίες.
- Το ότι μπορούν να ληφθούν υπόψη ή και να εξαιρεθούν ακραίες τιμές για διαφορετικές παραμέτρους που ενδεχομένως να αποτελούν περιοριστικούς παράγοντες για τη λήψη βιολογικώς έγκυρων αποτελεσμάτων.
- Το γεγονός ότι με τη χρήση των μαθηματικών μοντέλων μπορούμε να λάβουμε απαντήσεις σε «what-if?»ερωτήματα.

Κατόπιν των ανωτέρω γίνεται εύκολα αντιληπτό πως η αξιοποίηση των δυνατοτήτων των in silico μεθόδων συνδυαστικά με τις προϋπάρχουσες in vivo και in vitro μπορεί να δώσει αναρίθμητες δυνατότητες. Δεν είναι τυχαίο άλλωστε το γεγονός πως η «αλυσίδα in vivo - in vitro - in silico» (*in silico – in vitro - in vivo chain*) χρήζει τα τελευταία χρόνια όλο και ευρύτερης αποδοχής από επιστήμονες διαφόρων ειδικοτήτων λόγω της μεγάλης προόδου που σημειώθηκε τα τελευταία χρόνια. Τα μοντέλα in silico λειτουργούν ως πολύτιμες πηγές εισόδου δεδομένων για in vitro και in vivo πειράματα (συνεχές βέλος) και αντίστροφα οι in vitro και in vivo τεχνικές προσφέρουν ανατροφοδότηση στα in silico μοντέλα για την υποστήριξη περαιτέρω εξελίξεων και βελτιστοποίησης (διακεκομμένο βέλος)[16] όπως φαίνεται και στην παρακάτω εικόνα.



Εικόνα 12: The in silico-in vitro-in vivo chain
Πηγή: Marcu, L. G., & Harriss-Phillips, W. M. (2012))

3.2. Μοντέλα QSAR

Μια κατηγορία των in silico μεθόδων που χρησιμοποιείται ευρύτατα τελευταία είναι τα μοντέλα που συσχετίζουν τη δομή με την δραστικότητα (Structure Activity Relationships, SAR) και τα μοντέλα ποσοτικής σχέσης δομής – δραστικότητας (Quantitative Structure Activity Relationships, QSAR). Σύμφωνα με τον Ευρωπαϊκό Οργανισμό Χημικών Προϊόντων (ECHA) πρόκειται για μαθηματικά μοντέλα που μπορούν να χρησιμοποιηθούν για την πρόβλεψη των φυσικοχημικών και βιολογικών ιδιοτήτων των ενώσεων, με βάση τη χημική δομή τους. Κοινώς θα μπορούσαμε να πούμε πως ένα QSAR μοντέλο περιγράφεται από την παρακάτω εξίσωση:

Βιολογική Δράση = f (Δομή)

Οι ποσοτικές σχέσεις δομής – δραστικότητας βασίζονται αφενός στην υπόθεση ότι η δομή ενός μορίου σχετίζεται με τα χαρακτηριστικά εκείνα που είναι υπεύθυνα για τις φυσικές, χημικές ή βιολογικές ιδιότητες και αφετέρου στην δυνατότητα προσδιορισμού μιας ουσίας μέσω μιας ή περισσοτέρων αριθμητικών παραμέτρων. Με την χρήση των σχέσεων-μοντέλων QSAR καθίσταται δυνατός ο προσδιορισμός της βιολογικής συμπεριφοράς, ιδιότητας ή δραστικότητας μιας νέας ουσίας με βάση τη μοριακή δομή μιας άλλης, παρόμοιας, ουσίας, της οποίας η αντίστοιχη ιδιότητα έχει ήδη εκτιμηθεί. Όταν μιλάμε για προσδιορισμό ιδιότητας μιλάμε για σχέση QSPR (ποσοτική δράση δομής - ιδιότητας).

3.2.1. Ιστορική Αναδρομή

Οι ποσοτικές σχέσεις δομής – δραστικότητας βασίζονται αφενός στην υπόθεση ότι η δομή ενός μορίου σχετίζεται με τα χαρακτηριστικά εκείνα που είναι υπεύθυνα για τις φυσικές, χημικές ή βιολογικές ιδιότητες και αφετέρου στην δυνατότητα προσδιορισμού μιας ουσίας μέσω μιας ή περισσοτέρων αριθμητικών παραμέτρων. Με την χρήση των σχέσεων-μοντέλων QSAR καθίσταται δυνατός ο προσδιορισμός της βιολογικής συμπεριφοράς, ιδιότητας ή δραστικότητας μιας νέας ουσίας με βάση τη μοριακή δομή μιας άλλης, παρόμοιας, ουσίας, της οποίας η αντίστοιχη ιδιότητα έχει ήδη εκτιμηθεί. Όταν μιλάμε για προσδιορισμό ιδιότητας μιλάμε για σχέση QSPR (ποσοτική δράση δομής - ιδιότητας).

Αν και η χρήση των ποσοτικών σχέσεων δομής – δραστικότητας παρουσιάζει τρομερή άνθιση τις τελευταίες δεκαετίες, προφανώς λόγω της επιστημονικής εξέλιξης στους τομείς της πληροφορικής και της τεχνολογίας που παρέχει τη δυνατότητα σκιαγράφησης και φιλτραρίσματος των πολλαπλών μεταβλητών που χρησιμοποιούνται στην μοντελοποίηση επιτυγχάνοντας παράλληλα τους στόχους του περιορισμού της πραγματοποίησης in vivo πειραμάτων και της σύνθεσης μικρότερου αριθμού ενώσεων, οι

πρώτες αναφορές στις σχέσεις QSAR βρίσκονται πολύ πίσω στο χρόνο, για την ακρίβεια πάνω από έναν αιώνα πριν.

Το 1863 ο Cros A.F.A., [6] κατά την εκπόνηση της διατριβής του στο Πανεπιστήμιο του Στρασβούργου, έκανε λόγο για την αντίστροφη σχέση που δίεπε την τοξικότητα των πρωτοταγών αλειφατικών αλκοολών και την διαλυτότητά τους στο νερό. Η σχέση αυτή ήταν σύμφωνη με το θεμελιώδες αξίωμα των σχέσεων δομής – τοξικότητας αφού επιβεβαίωνε τον συσχετισμό μεταξύ της χημικής δομής και της τοξικότητας καθώς η τοξικότητα ήταν απόρροια των ιδιοτήτων της ουσίας ως αποτέλεσμα της χημικής της δομής.

Πέντε χρόνια αργότερα, το 1868, οι Crum-Brown και Fraser κατά τη διάρκεια της έρευνάς τους σε διάφορα αλκαλοειδή διατύπωσαν την άποψη ότι η βιολογική δράση των χημικών ενώσεων είναι συνάρτηση της δομής τους διατυπώνοντας μάλιστα και την πρώτη εξίσωση μαθηματικοποίησης της ποσοτικής σχέσης δομής δραστικότητας η οποία είναι η ακόλουθη:

Βιολογική Δράση = f (Δομή)

Την ίδια περίπου περίοδο, το 1869, ο Ρώσος χημικός και εφευρέτης Dmitri Mendeleev, παρουσίασε στην Αγία Πετρούπολη μια πρώτη μορφή του γνωστού σήμερα Περιοδικού Πίνακα Χημικών Στοιχείων ο οποίος τότε περιείχε μόλις 63 στοιχεία αφήνοντας θέσεις για τα στοιχεία που θα ανακαλύπτονταν μελλοντικά. Ωστόσο έκανε πρόβλεψη για τρία χημικά στοιχεία τα οποία ανακαλύφθηκαν λίγα χρόνια αργότερα. Αυτά ήταν το Γάλλιο (Ga) που ανακαλύφθηκε το 1875 από τον Γάλλο χημικό Lecoq de Boisbaudran, το Σκάνδιο (Sc) που ανακαλύφθηκε το 1879 από τον Lars Fredrik Nilson και το Γερμάνιο (Ge) που ανακαλύφθηκε το 1886 από τον Clemens Winkler.[25] Το γεγονός πως ο Mendeleev δεν προέβλεψε μόνο την ανακάλυψη των τριών αυτών χημικών στοιχείων αλλά και κάποιες από τις ιδιότητές τους, βασιζόμενος στην θέση τους στον Περιοδικό Πίνακα, τον καθιστούν

μεταξύ των πρώτων επιστημόνων που συσχέτισαν την δομή με τη δραστικότητα των ουσιών.

Το 1890 ο Hans Horst Meyer από το Πανεπιστήμιο του Μάρμπουργκ της Γερμανίας και ο Charles Ernest Overton από το Πανεπιστήμιο της Ζυρίχης δουλεύοντας ανεξάρτητα παρατήρησαν ότι η τοξικότητα των οργανικών ενώσεων σχετιζόταν με την λιποφιλία τους.[28]

Τα αμέσως επόμενα χρόνια, και συγκεκριμένα το 1893, ο Richet έκανε την διαπίστωση πως η κυτταροτοξικότητα μιας ομάδας απλών οργανικών ενώσεων και η διαλυτότητά τους στο νερό είχαν σχέση αντιστρόφως ανάλογη.[26] Λίγο αργότερα, το 1899 ο Meyer [17] και το 1901 ο Overton,[23] συσχέτισαν την λιποφιλία (εκφρασμένης με το συντελεστή μερισμού στο σύστημα ελαίου-νερού) με τη γενική αναισθητική δράση.

Το 1939 ο Ferguson, διατύπωσε την άποψη πως η τάση ατμών των τοξικών συγκεντρώσεων (C_T) μιας σειράς ενώσεων θα μπορούσε να προβλεφθεί από την υδατοδιαλυτότητα τους ή την τάση ατμών (όταν επρόκειτο για πτητικές ουσίες) σύμφωνα με την παρακάτω εξίσωση:

$$\mathbf{C_T = kA^{1/n}}$$

Όπου: $\mathbf{C_T}$: η τοξική συγκέντρωση

k, n: σταθερές με n > 1

A: η διαλυτότητα ή η τάση ατμών

Λίγα χρόνια αργότερα, το 1948, ο Ferguson σε συνεργασία με τον Pirie παρουσίασε μια γενικευμένη μορφή της παραπάνω εξίσωσης όπου το k μπορούσε να είναι αντί για σταθερά μια άλλη φυσικοχημική ιδιότητα όπως π.χ. ο συντελεστής κατανομής. Η γενικευμένη αυτή μορφή αποτέλεσε τη βάση για τον Purcell και τους συνεργάτες τους που το 1973 παρατήρησαν ότι η εξίσωση μπορεί να μετασχηματιστεί λογαριθμικά στη μορφή που μέχρι και σήμερα έχουν τα μοντέλα QSAR όπως φαίνεται παρακάτω: [14]

$$\log 1/C = k' + n'.\log A$$

Το 1939 επίσης ο Louis Hammett έθεσε γερές βάσεις για τις σύγχρονες μεθόδους QSAR εισάγοντας τις ηλεκτρονιακές σταθερές και διατυπώνοντας την άποψη ότι αυτές εκφράζουν την επίδραση που ασκούν οι υποκαταστάτες των οργανικών ενώσεων στο ρυθμό πραγματοποίησης μιας χημικής αντίδρασης. Σύμφωνα με την θεωρία του Hammett η επίδραση αυτή είναι σταθερή αφενός για κάθε υποκαταστάτη και αφετέρου για κάθε αντίδραση. Κάποια χρόνια αργότερα, και συγκεκριμένα την δεκαετία του '50 η εξίσωση του Hammett τροποποιήθηκε με την εισαγωγή της στερικής σταθεράς υποκαταστάτη, Es, ύστερα από πρόταση του Taft, λαμβάνοντας επιπλέον υπόψη και τα στερικά φαινόμενα που προκαλλούνται από τους υποκαταστάτες και ασκούν επιρροή στον ρυθμό μιας χημικής αντίδρασης.

Η προσέγγιση των Hammett και Taft αποτέλεσε εφαλτήριο για τους ερευνητές και έτσι λίγο αργότερα, το 1964, δημιουργήθηκε το πρώτο μοντέλο Ποσοτικών Σχέσεων Δομής-Δράσης, ύστερα από την διαπίστωση των Hansch και Fujita πως η εξίσωση Hammett – Taft θα μπορούσε να χρησιμοποιηθεί για να περιγράψει εκτός των άλλων και την βιολογική δράση. Κατέληξαν λοιπόν στην παρακάτω εξίσωση η οποία λαμβάνει ως δεδομένο ότι η βιολογική δράση είναι συνάρτηση τριών κατηγοριών ιδιοτήτων: των στερικών ιδιοτήτων, των ηλεκτρονιακών ιδιοτήτων και της λιποφιλίας. [33]

Βιολογική Δράση=f(λιποφιλία + ηλεκτρονιακές ιδιότητες + στερικές ιδιότητες)

Και πιο συγκεκριμένα:

$$\text{Log } 1/C50 = -(logP)2 + blogP + \rho\sigma + \delta Es + c$$

Όπου: Log 1/C50 : έκφραση του λογαρίθμου της βιολογικής απόκρισης

logP : σταθερά μερισμού ως μέτρο της λιποφιλίας

σ : η ηλεκτρονιακή σταθερά του Hammett

Es : η στερική σταθερά του Taft

a, b, ρ, δ: συντελεστές που προκύπτουν από πολλαπλή γραμμική ανάλυση παλινδρόμησης

c : σταθερός όρος που προκύπτει από πολλαπλή γραμμική ανάλυση παλινδρόμησης

Την ίδια περίοδο, το 1964, πραγματοποιήθηκε μια ακόμη προσέγγιση στην ποσοτική περιγραφή των σχέσεων δομής δραστηριότητας από τους Free και Wilson. Σε αντίθεση με την κατά Hansch προσέγγιση που συνέδεε τις φυσικοχημικές ιδιότητες με τις τιμές της βιολογικής δραστηριότητας, η προσέγγιση των Free και Wilson συνέδεε τα δομικά χαρακτηριστικά με τις βιολογικές ιδιότητες σύμφωνα με την παρακάτω εξίσωση:

$$log1/C=\mu_o + \Sigma\alpha_{ik}x_{ik}$$

Όπου α: η συνεισφορά στη βιολογική δράση του υποκαταστάτη x στη θέση i στο μόριο k

μο: η μέση δραστικότητα

Αν και φαινομενικά οι δύο προσεγγίσεις δεν είχαν κοινά χαρακτηριστικά επί της ουσίας είναι στενά αλληλένδετες, τόσο σε θεωρητικό επίπεδο όσο και στην πρακτική εφαρμογή τους αφού συχνά καθίσταται δυνατή η συνδυαστική χρήση των δύο μοντέλων και η χρήση παραμέτρων Free - Wilson για την περιγραφή της επίδραση των υποκαταστατών στη βιολογική δραστηριότητα με τον μεγάλο αριθμό επιτυχημένων εφαρμογών να καταδεικνύει ότι το συνδυαστικό αυτό μοντέλο είναι ένα από τα ισχυρότερα εργαλεία της κλασικής μεθόδου QSAR.[13]

Τις δεκαετίες που ακολούθησαν, πραγματοποιήθηκε μεγάλη εξέλιξη στην ανάπτυξη των QSAR τεχνικών, τις δεκαετίες του 1970 και το 1980 υπήρξε ανάπτυξη Ποσοτικών Σχέσεων Δομής - Δράσης δύο διαστάσεων (2D - QSAR, Two Dimensional Quantitative Structure - Activity Relationship) ενώ το 1988 αποτελεί ορόσημο αφού ο R. Cramer εισήγαγε για πρώτη φορά τις Ποσοτικές Σχέσεις δομής - δράσης τριών διαστάσεων (3D - QSAR, Three Dimensional Quantitative Structure - Activity Relationship).[34] Από τότε και

έπειτα με τη βοήθεια της τεχνολογικής εξέλιξης που είναι ραγδαία στην εποχή μας οι τεχνικές QSAR κερδίζουν όλο και περισσότερο έδαφος στην επιστημονική κοινότητα και έρευνα αφού βρίσκουν εφαρμογές σε πολλούς κλάδους της επιστήμης με κυριότερο αυτόν του σχεδιασμού φαρμάκων.

3.2.2. Βήματα δημιουργίας μοντέλων QSAR

Για την δημιουργία ενός ποιοτικού μοντέλου ποσοτικής σχέσης δομής-δραστικότητας απαιτούνται τα κάτωθι βασικά βήματα:

- **<u>Βήμα 1ο</u>: Η επιλογή των προς μελέτη ενώσεων/δεδομένων**

  Θα πρέπει να επιλεγούν δεδομένα που περιέχουν καλά κατανεμημένες τιμές βιολογικής δράσης ως προς τον στόχο (μεταβλητή απόκρισης) και να διαθέτουν επαναληψιμότητα. Μετά την επιλογή τους τα δεδομένα διαχωρίζονται σε δύο αντιπροσωπευτικά σύνολα δεδομένων. Το ένα χρησιμοποιείται για την εκπαίδευση του μοντέλου (training dataset) και το άλλο για την επαλήθευση του μοντέλου (test dataset). Για τον διαχωρισμό αυτό υπάρχουν διαθέσιμες αρκετές μέθοδοι της στατιστικής.

- **<u>Βήμα 2ο</u>: Η επιλογή των παραμέτρων**

  Από όλες τις διαθέσιμες παραμέτρους που διαθέτουμε για τις υπό εξέταση ενώσεις επιλέγονται εκείνες με την μεγαλύτερη συνεισφορά στον υπολογισμό της παραμέτρου στόχου που οδηγούν στην δημιουργία ενός αξιόπιστου μοντέλου.

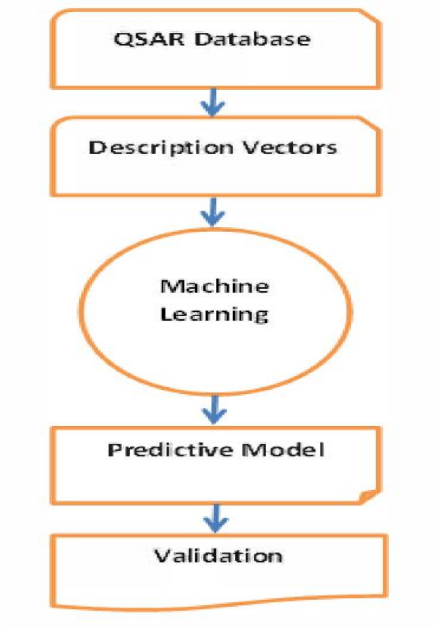
- **<u>Βήμα 3ο</u>: Η επιλογή της μεθόδου, αλγορίθμου**

  Επιλέγεται ο κατάλληλος αλγόριθμος για την μοντελοποίηση της ποσοτικής σχέσης δομής-δραστικότητας. Η επιλογή γίνεται ανάλογα με την υπό εξέταση περίπτωση και τα υπό εξέταση δεδομένα καθώς κάθε αλγόριθμος παρέχει διαφορετικές δυνατότητες και ανταποκρίνεται διαφορετικά σε διαφορετικές περιπτώσεις δεδομένων.

Οι βασικότεροι αλγόριθμοι της Μηχανικής Μάθησης που χρησιμοποιούνται στην δημιουργία μοντέλων QSAR παρουσιάστηκαν αναλυτικά στο Κεφάλαιο 2.

- **<u>Βήμα 4ο</u>: Ο έλεγχος αξιοπιστίας και ερμηνεία του μοντέλου**

Επειδή δεν αρκεί μόνο η δημιουργία του μοντέλου αλλά πρέπει να εξασφαλίσουμε ότι οι προβλέψεις που λαμβάνουμε είναι αξιόπιστες, το τελευταίο και βασικότερο βήμα είναι ο έλεγχος αξιοπιστίας και η επικύρωση του μοντέλου QSAR που προέκυψε αλλά και η ερμηνεία του σε συνδυασμό με τον καθορισμό του πεδίου εφαρμογής του.



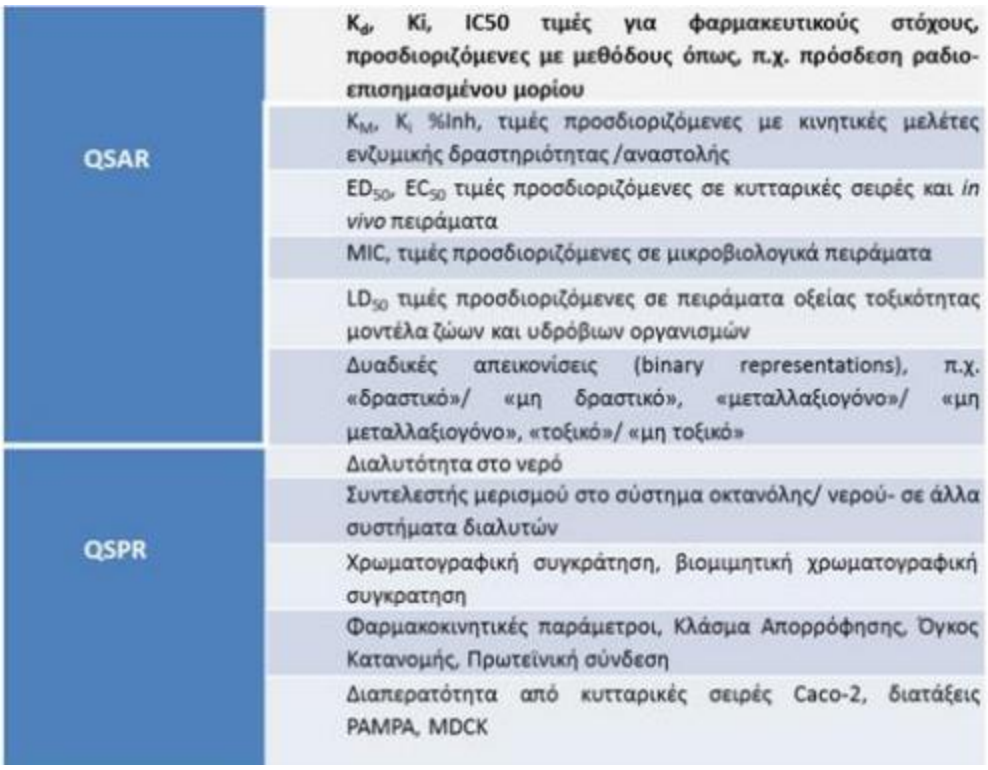
Εικόνα 13: General Steps of Developing QSAR Models
Πηγή: https://www.semanticscholar.org/

3.2.3. Μεταβλητές απόκρισης (Endpoints)

Όπως αναφέρθηκε παραπάνω το πρώτο βήμα στην διαδικασία εξαγωγής μοντέλων QSAR είναι η επιλογή των προς μελέτη ενώσεων και ο καθορισμός της μεταβλητής απόκρισης (endpoint). Για τον λόγο αυτό τα υπό μελέτη βιολογικά δεδομένα πρέπει να πληρούν κάποιες προϋποθέσεις. Σύμφωνα με τους Δημόπουλο και Τσαντίλη-Κακουλίδου στις προϋποθέσεις αυτές συγκαταλέγονται τα εξής: Πρώτα απ' όλα θα πρέπει όλες οι υπό μελέτη ενώσεις να έχουν το ίδιο μηχανισμό δράσης και να δρουν στον ίδιο υποδοχέα. Η

βιολογική τους δράση θα πρέπει να εκφράζεται σε αριθμητικά δεδομένα, όταν πρόκειται να χρησιμοποιηθούν σε πρόβλημα regression (παλινδρόμησης) ή σε μη αριθμητικά δεδομένα που όμως τις διαφοροποιούν ανάλογα με την δράση τους όταν πρόκειται να χρησιμοποιηθούν σε προβλήματα classification (ταξινόμησης). Επιπρόσθετα θα πρέπει η βιολογική δράση να κατανέμεται ορθώς, να είναι διαφοροποιημένη και να υπάρχουν διαθέσιμα δεδομένα για όλο το εύρος της. Τα δεδομένα αυτά προκύπτουν από βιολογικά πειράματα για τα οποία θα πρέπει να είναι γνωστές οι πληροφορίες αφενός σχετικά με το επίπεδο διεξαγωγής τους (εάν είναι π.χ. σε μοριακό επίπεδο, σε κυτταρικό επίπεδο, σε όργανο, σύστημα ή οργανισμό, ή αν πρόκειται για πειράματα in situ) και αφετέρου με την αξιοπιστία και την επαναληψιμότητά τους. Σε αυτό το σημείο να σημειωθεί πως τα στατιστικά στοιχεία του μοντέλου δεν θα πρέπει να είναι καλύτερα από αυτά των βιολογικών πειραμάτων.

Στην παρακάτω εικόνα εμφανίζονται τα φυσικοχημικά και βιολογικά μεγέθη που χρησιμοποιούνται συνηθέστερα ως μεταβλητές απόκρισης σε μελέτες QSAR και QSPR.



Εικόνα 14: Φυσικοχημικά και βιολογικά μεγέθη που χρησιμοποιούνται ως μεταβλητές απόκρισης σε μελέτες QSAR και QSPR.

Πηγή: Δημόπουλος, Β., Τσαντίλη-Κακουλίδου, Ά. 2015. Βασικές αρχές σχεδιασμού και ανάπτυξης φαρμάκων

3.2.4. Περιγραφικές Μεταβλητές (Descriptors)

Σύμφωνα με τους Todeschini & Consonni η περιγραφική μεταβλητή είναι μια μαθηματική αναπαράσταση της χημικής δομής και ορίζεται ως: *«Μια περιγραφική μεταβλητή είναι το τελικό αποτέλεσμα μιας λογικής και μαθηματικής διαδικασίας που μετατρέπει χημικές πληροφορίες που κωδικοποιούνται μέσα σε μια συμβολική αναπαράσταση ενός μορίου σε έναν χρήσιμο αριθμό ή το αποτέλεσμα κάποιου τυποποιημένου πειράματος»*.[30]

Υπάρχουν χιλιάδες περιγραφικές μεταβλητές που χρησιμοποιούνται στις μελέτες QSAR και QSPR. Οι βασικές κατηγορίες είναι οι παρακάτω: [9]

- **Μηδενικής διάστασης (0-D):** Προέρχονται αποκλειστικά και μόνο από τη δομή και περιέχουν πληροφορίες σχετικά με την σύνθεση π.χ. αριθμός ατόμων Ν.
- **1-D /αποτυπώματα:** Περιγράφουν τη σύνθεση σε όρους δομικών θραυσμάτων, π.χ. αριθμός βενζοϊκών δακτυλίων.
- **2-D / τοπολογικές:** Περιλαμβάνουν πληροφορίες σχετικά με τη συνδεσιμότητα των ατόμων και τα δομικά θραύσματα. Λαμβάνονται από την 2D αναπαράσταση της χημικής δομής και έχουν την δυνατότητα διαφοροποίησης των μορίων βασιζόμενες σε ιδιότητες όπως το μέγεθός, ο βαθμός διακλάδωσης, η ευελιξία, κλπ.
- **3-D / γεωμετρικές**: Εξαρτώνται από την 3D αναπαράσταση της χημικής δομής και περιλαμβάνουν πληροφορίες σχετικά με το μέγεθος, το σχήμα, την επιφάνεια, τον όγκο και ηλεκτρονιακά (κβαντικά) στοιχεία.
- **4-D**: Είναι παρόμοιες με τις 3-D, αλλά αποτυπώνουν επιπλέον τις ενέργειες αλληλεπίδρασης.
- **Φυσικοχημικές ιδιότητες:** Μπορούν επίσης να χρησιμοποιηθούν ως περιγραφικές μεταβλητές αλλά χρησιμοποιούνται περισσότερο ως μεταβλητές απόκρισης. Η ιδιότητα μπορεί να υπολογιστεί τόσο από την χημική δομή όσο και από πειραματική μέτρηση. Ένα κοινό παράδειγμα φυσικοχημικής ιδιότητας που χρησιμοποιείται συχνά ως endpoint σε μελέτες QSAR είναι το logP.

3.2.5. Ελεγχος αξιοπιστίας (validation)

Όπως αναφέρθηκε ήδη στην παράγραφο 3.2.2. το βασικότερο στάδιο στις μελέτες QSAR είναι η επικύρωση της αξιοπιστίας του παραγόμενου μοντέλου. Οι τεχνικές QSAR, κάνοντας χρήση στατιστικών εργαλείων, παράγουν προγνωστικά μοντέλα που συσχετίζουν περιγραφικές μεταβλητές αντιπροσωπευτικές της της μοριακής δομής ή ιδιοτήτων με την βιολογική συμπεριφορά. Η παραγωγή ενός ποιοτικού μοντέλου QSAR είναι συνάρτηση πολλών παραγόντων όπως η ποιότητα των δεδομένων εισόδου, η ορθή επιλογή περιγραφικών μεταβλητών και η σωστή επιλογή αλγορίθμων για την μοντελοποίηση και την επικύρωση. Το γεγονός ότι τα μοντέλα αυτά χρησιμοποιούνται σε ιδιαίτερα ευρύ φάσμα κομβικών επιστημονικών τομέων καθιστά την παραγωγή ενός αξιόπιστου μοντέλου απαραίτητη, γίνεται λοιπόν αντιληπτό ότι οποιοδήποτε μοντελοποίηση QSAR θα πρέπει να οδηγεί σε μοντέλα ικανά να κάνουν ακριβείς και αξιόπιστες προβλέψεις.

Για την επικύρωση των μοντέλων QSAR χρησιμοποιούνται διάφορες μέθοδοι με συνηθέστερες την εξωτερική και την εσωτερική επικύρωση.[31] Η εσωτερική επικύρωση στοχεύει στην διαπίστωση ότι το μοντέλο περιγράφει ικανοποιητικά τα δεδομένα από τα οποία δημιουργήθηκε ενώ η εξωτερική στην διαπίστωση ότι το μοντέλο δύναται να περιγράψει ικανοποιητικά δεδομένα που δεν έχουν χρησιμοποιηθεί για την δημιουργία του.

Εσωτερική επικύρωση αποτελεί η διασταυρούμενη επικύρωση (cross validation) κατά την οποία η ανάλυση επαναλαμβάνεται πολλές φορές εξαιρόντας κάθε φορά μια ένωση (leave one out) ή ομάδα ενώσεων (k-fold cross validation) από το σύνολο δεδομένων εως ότου εξασφαλιστεί ότι όλες οι ενώσεις έχουν αποκλειστεί εκ περιτροπής.

Στην περίπτωση του leave one out cross validation για σύνολο δεδομένων που περιέχει (ν) δείγματα θα πραγματοποιηθούν ν πειράματα (επαναλήψεις) σε καθένα από τα οποία (ν-1) πλήθος δειγμάτων θα χρησιμοποιείται ως training data και 1 ως test data. Το

συνολικό ποσοστό λάθος του μοντέλου ισούται με τον μέσο όρο των λαθών των ν επαναλήψεων σύμφωνα με την σχέση $\mathrm{E} = \frac{1}{\nu}\sum_{i=1}^{\nu}\mathrm{E}_i$

Η περίπτωση k-fold cross validation αποτελεί γενίκευση του leave one out. Ειδικότερα το αρχικό δείγμα χωρίζεται τυχαία σε k υπο-δείγματα ίσου μεγέθους και πραγματοποιούνται k επαναλήψεις της διαδικασίας επικύρωσης. Σε κάθε επανάληψη (k-1) πλήθος υπο-δειγμάτων χρησιμοποιείται ως training data και 1 υπο-δείγμα ως test data. Η διαδικασία διασταυρούμενης επικύρωσης επαναλαμβάνεται k φορές, ώστε να εξασφαλιστεί πως καθένα από τα k υποδείγματα να χρησιμοποιείται ακριβώς μία φορά ως test data. Το συνολικό ποσοστό λάθος του μοντέλου ισούται και σε αυτήν την περίπτωση με τον μέσο όρο των λαθών των k επαναλήψεων σύμφωνα με την σχέση $\mathrm{E} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{E}_i$

Παραλλαγή αυτής της μεθόδου θα μπορούσε να θεωρηθεί η επανειλημμένη τυχαία δειγματοληψία γνωστή και ως διασταυρούμενη επικύρωση Monte Carlo,[7] η οποία δημιουργεί τυχαίους διαχωρισμούς του συνόλου δεδομένων σε training data και test data και επιδιώκει την επικύρωση του μοντέλου με χρήση των δεύτερων. Για κάθε τέτοια διάσπαση, το μοντέλο είναι κατάλληλο για τα δεδομένα εκπαίδευσης και η προβλεπτική ακρίβεια αξιολογείται χρησιμοποιώντας τα δεδομένα επικύρωσης. Το συνολικό ποσοστό λάθος του μοντέλου ισούται και σε αυτήν την περίπτωση από τον μέσο όρο. Αν και το γεγονός ότι το ποσοστό της επικύρωσης δεν εξαρτάται από τον αριθμό των επαναλήψεων αποτελεί σημαντικό πλεονέκτημα έναντι της επικύρωσης k-fold cross εντούτοις αποτελεί πρόβλημα το ότι ορισμένες παρατηρήσεις ενδεχομένως να μην επιλεγούν ποτέ στο υποσύνολο επικύρωσης, ενώ άλλες πιθανόν να επιλεχθούν περισσότερες από μία φορές με αποτέλεσμα τα υποσύνολα ενδέχεται να επικαλύπτονται. Σε αντίθεση την μέθοδο k-fold cross validation όπου όλες οι παρατηρήσεις χρησιμοποιούνται τόσο για την εκπαίδευση όσο και για την επικύρωση και κάθε παρατήρηση χρησιμοποιείται test data ακριβώς μία φορά. Επιπρόσθετα το ότι η μέθοδος παρουσιάζει επίσης μεταβλητότητα Monte Carlo, υποδηλώνει ότι τα αποτελέσματα θα διαφέρουν εάν η ανάλυση επαναληφθεί με διαφορετικά τυχαία υποσύνολα δεδομένων.

4. Περιγραφή Εργαλείων & Μοντέλων που υλοποιήθηκαν

4.1. Γενικά

Έχει αναφερθεί από το πρώτο κεφάλαιο ακόμη πως τα κατασκευασμένα νανοϋλικά (ENM) κερδίζουν όλο και περισσότερο έδαφος στα περισσότερα πεδία της έρευνας και της επιστήμης. Σημαντικό ωστόσο βήμα για την ορθή αξιοποίηση των πλεονεκτημάτων που διαθέτουν είναι η κατανόηση και η αξιολόγηση των μηχανισμών τοξικότητάς τους. περισσότερο τις ζωές μας λόγω των εφαρμογών τους σε πολλά πεδία. Όπως αναφέρεται και στο κεφάλαιο 3 τα τελευταία χρόνια ακολουθώντας τις εξελίξεις στον τομέα της επιστήμης των υπολογιστών έχουν κερδίσει ιδιαίτερο έδαφος οι in silico μέθοδοι οι οποίες τείνουν σε μεγάλο βαθμό να αντικαταστήσουν τις παραδοσιακές πειραματικές μεθόδους αφού είναι πιο συμφέρουσες οικονομικά, λιγότερο χρονοβόρες και παράλληλα συμμορφώνονται με την υπ' αριθμ. 2010/63/EU Οδηγία του Ευρωπαϊκού Κοινοβουλίου σχετικά με την μείωση των εργαστηριακών δοκιμών σε ζώα. [1]

Κινούμενο προς αυτή την κατεύθυνση η ομάδα του Εργαστηρίου Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου, υπό την καθοδήγηση του Καθηγητή κ. Χαράλαμπου Σαρίμβει, στα πλαίσια του χρηματοδοτούμενου Ευρωπαϊκού Πρόγραμματος NanoCommons, δημιούργησε δύο εκτεταμένες εφαρμογές ανοιχτού κώδικα στο διαδίκτυο που παρέχουν δυνατότητες nanoQSAR μοντελοποίησης με σκοπό την πρόβλεψη των αρνητικών επιπτώσεων των νανοϋλικών. Πρόκειται για τις εφαρμογές Jaqpot Quattro και Jaqpot v5 που αποτελεί την εξέλιξη του Jaqpot Quattro και παρουσιάζονται παρακάτω:

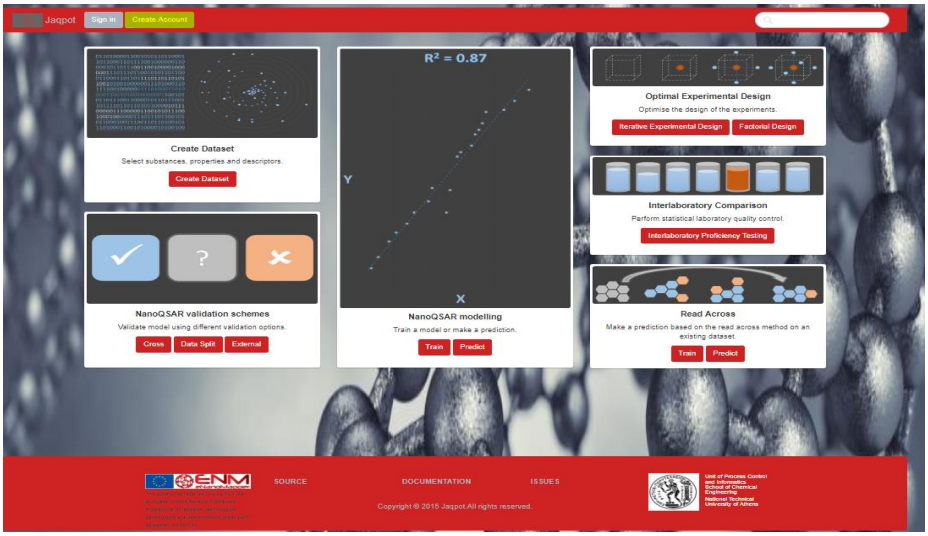
4.2. Jaqpot Quattro [5]

Η εφαρμογή ανοιχτού κώδικα Jaqpot Quattro είναι διαθέσιμη στην ηλεκτρονική διεύθυνση http://www.jaqpot.org/ και παρέχει στον χρήστη πληθώρα δυνατοτήτων όπως αυτές της εισαγωγής, επιλογής και επεξεργασίας δεδομένων που διατίθενται στην βάση δεδομένων eNanoMapper. Επιπλέον διαθέτει επιλογές προετοιμασίας και προεπεξεργασίας δεδομένων ώστε να χρησιμοποιηθούν στην μοντελοποίηση, αλλά και υπηρεσίες υπολογισμού περιγραφικών μεταβλητών από πρωτογενή δεδομένα (π.χ. εικόνες) ενώ παρέχει και τη δυνατότητα χρήσης PMML. Το Jaqpot Quattro ενσωματώνει αλγορίθμους στατιστικής και εξόρυξης δεδομένων με ταυτόχρονη δυνατότητα του χρήστη να δημιουργήσει τους δικούς του για την δημιουργία μοντέλων σχέσεων νανο-ποσοτικής δομής-δραστηριότητας (nanoQSAR modeling). Τα παραγόμενα μοντέλα δύνανται στην συνέχεια να επικυρωθούν μέσω των παρεχόμενων από το σύστημα δυνατοτήτων επικύρωσης (split-, cross- and external validation), ενώ υποστηρίζεται και η παραγωγή επεξεργάσιμων αναφορών QPRF (QSAR Prediction Reporting Format) οι οποίες μπορούν να αποθηκευτούν σε μορφή pdf. Η εφαρμογή διαθέτει επίσης ένα αποθετήριο δεδομένων αλλά και υλοποιημένων μοντέλων της βιβλιογραφίας τα οποία μπορεί ο χρήστης να χρησιμοποιήσει για να κάνει προβλέψεις βασιζόμενος στα δικά του δεδομένα ενώ αν θέλει να αποθηκεύσει τα δεδομένα του ή τα μοντέλα που ο ίδιος δημιούργησε ώστε να του είναι προσβάσιμα από κάθε συσκευή με δυνατότητα σύνδεσης στο διαδίκτυο μπορεί να το κάνει στον ιδιωτικό χώρο που του παρέχεται μετά την δημιουργία του λογαριασμού του.

Καθώς έχει ήδη γίνει αναφορά στην διαδικασία που ακολουθείται για την δημιουργία ενός μοντέλου QSAR παρακάτω παρουσιάζονται και οπτικοποιούνται τα βήματα στη εφαρμογή Jaqpot Quattro.

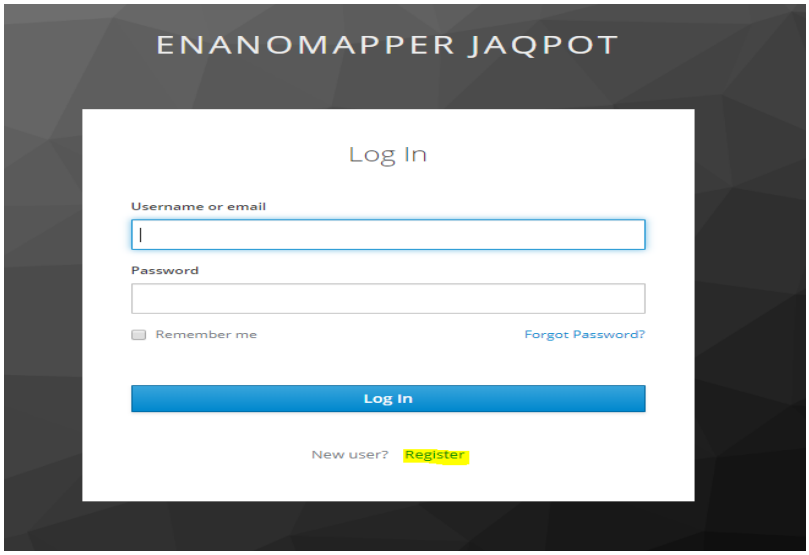
4.2.1. Σύνδεση/Εγγραφή & Αρχικές οθόνες

Η εφαρμογή είναι διαθέσιμη στην ηλεκτρονική διεύθυνση: http://www.jaqpot.org/ και η πρώτη εικόνα είναι η παρακάτω:



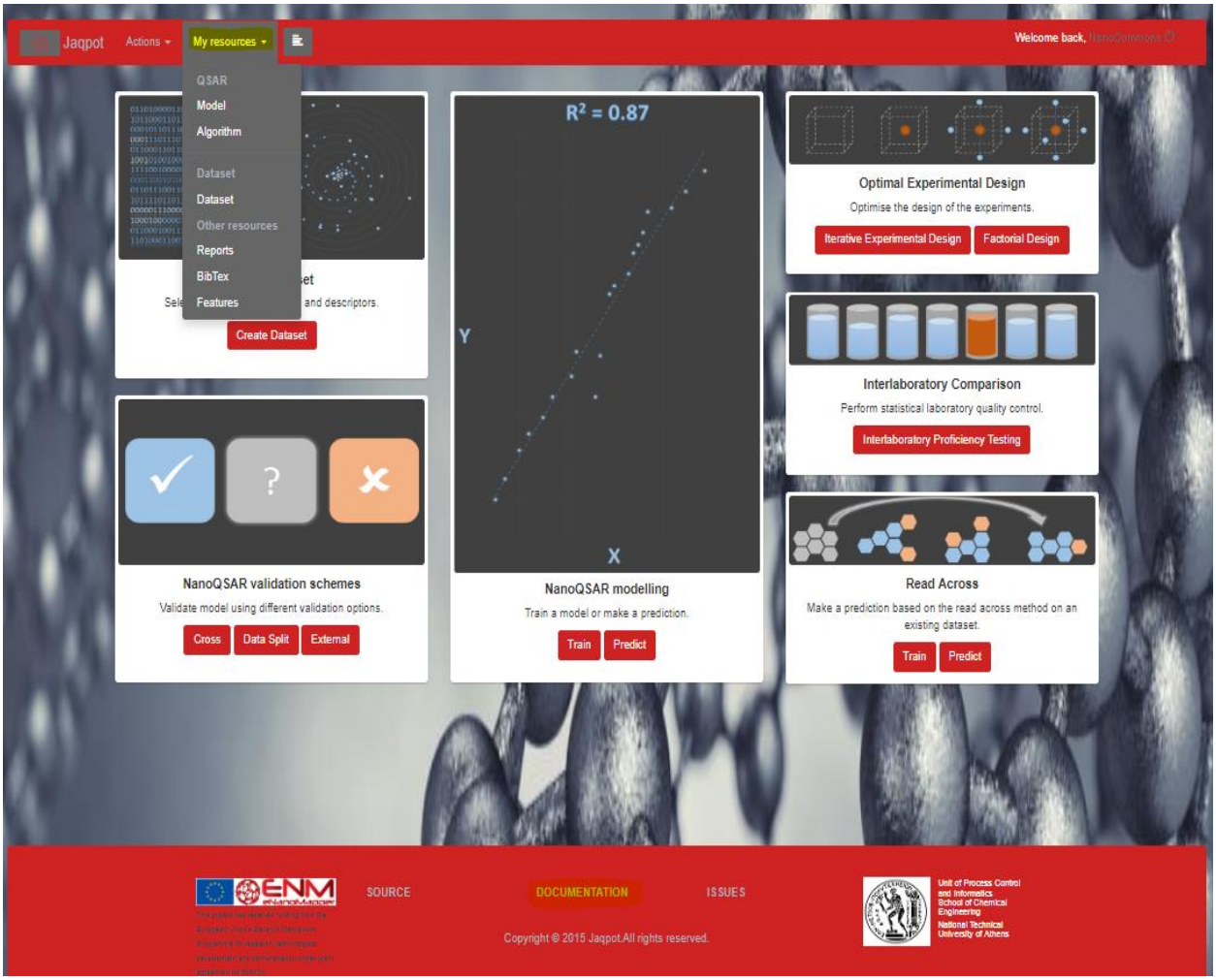
Εικόνα 15: Αρχική Οθόνη

Για την χρήση της πλατφόρμας απαιτείται η δημιουργία λογαριασμού, επιλέγοντας την επιλογή Create account, πάνω αριστερά ο χρήστης μεταφέρεται στην επόμενη σελίδα όπου πρέπει να επιλέξει Register



Εικόνα 16: Σελίδα Σύνδεσης

Μετά την δημιουργία του λογαριασμού ο χρήστης έχει πρόσβαση σε όλες τις λειτουργίες του εργαλείου αλλά και στο προσωπικό του αποθετήριο επιλέγοντας το «My resources» σύμφωνα με την εικόνα 17. Στις εικόνες 18 και 19 παρουσιάζεται το αποθετήριο των δεδομένων και μοντέλων αντίστοιχα. Στη άνω περιοχή (Example) φαίνονται όσα προϋπάρχουν στην εφαρμογή και ακριβώς από κάτω αυτά του χρήστη.



Εικόνα 17: Αρχική Εικόνα μετά την σύνδεση και πρόσβαση στο προσωπικό αποθετήριο του χρήστη.

Dataset:

Example datasets:

Name	Title	Description	Date
Gajewicz_10_29	Gajewicz et al - 10 Metal Oxide NPs	10 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.	2018-07-05T15:05:49.675+0000
Gajewicz_10_29_class	Gajewicz et al - 10 Metal Oxide NPs	10 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.	2018-07-05T15:05:49.959+0000
Gajewicz_18_29	Gajewicz et al - 18 Metal Oxide NPs	18 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.	2018-07-05T15:05:50.083+0000
Gajewicz_18_29_class	Gajewicz et al - 18 Metal Oxide NPs	18 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.	2018-07-05T15:05:50.192+0000
Gajewicz_8_29	Gajewicz et al - 8 Metal Oxide NPs	8 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.	2018-07-05T15:05:50.237+0000
Gajewicz_8_29_class	Gajewicz et al - 8 Metal Oxide NPs	8 MeOx NPs with 29 descriptors, used for predicting HaCaT toxicity.	2018-07-05T15:05:50.277+0000

All Datasets:

Name	Title	Description	Date
XmCQVC7o5jKKRv	Solubility of C60 fullerene in various solvents (tr...	The dataset includes 5 descriptors and solubility (log of molar fractions) for 93 solvents	2019-05-14T11:11:03.244+0000
nGF3G5SBo4wk5h	Solubility of C60 fullerene in various solvents	The dataset includes 5 descriptors and solubility (log of molar fractions) for 124 solvents	2019-05-14T11:09:50.746+0000
SUfmS4t4LoBREREdx5cp	Gharagheizi et al, Fullerene C60	Numerical values of the calculated descriptors along with solubility of C60 in solvents	2019-05-14T10:52:15.395+0000
xlYt2DN1NUct2G	Kar et al, metal oxides (training), mod_7	Numerical values of the calculated descriptors along with cytotoxicity values for metal oxidenanop...	2019-05-14T09:23:38.585+0000
FZxgQrEp2RBSWN	Marcus et al, Fullerene C60, all solvents, 298_303	Fullerene C60 Solubility (all solvents)	2019-05-06T11:02:02.822+0000
mZjbdhgfh1O7oi	Pathakoti et al, metal oxides, light (training)	Molecular properties of metal oxides	2019-05-06T08:49:42.616+0000
4xoqetJXfkMB0S	Pathakoti et al, metal oxides, light (full)	Molecular properties of metal oxides	2019-05-06T08:49:13.215+0000
r2Chi6fGaaHCrS	Pathakoti et al, metal oxides, dark (training)	Molecular properties of metal oxides	2019-05-06T08:48:55.173+0000
yjBO04fO3d19XL	Pathakoti et al, metal oxides, dark (full)	Molecular properties of metal oxides	2019-05-06T08:48:34.077+0000
llmv3I4960VK8n	Toporov et al, 2007, Fullerene C60, training	"Descriptors and values of the solubility, log S, of fullerene C60 in organic solvents"	2019-05-06T06:21:30.670+0000
u2kjmMsDbtpLwx	Pan et al, Metal oxides, full, mod_2	Descriptors and toxicity data for metal oxides	2019-05-05T23:01:38.921+0000
pK1E3Azl0s9vHA	Pan et al, Metal oxides, full, mod_1	Descriptors and toxicity data for metal oxides	2019-05-05T23:01:31.443+0000
2WPTL9B2z7XMYz	Pan et al, Metal oxides, training, mod_2	Descriptors and toxicity data for metal oxides	2019-05-05T22:59:45.106+0000
iUDhjXUvM5f3D3	Pan et al, Metal oxides, training, mod_1	Descriptors and toxicity data for metal oxides	2019-05-05T22:55:22.854+0000

Εικόνα 18: Αποθετήριο δεδομένων

Models

Example models:

Name	Title	Description	Date
gaj-10-linear	Gajewicz et al - QSAR Linear Regression on 10...	Linear Regression using 10 MeOx NPs, with selected descriptors (normalised): Standard Enthalpy...	2016-03-15T14:19:17.452+0000
gaj-18-linear	Gajewicz et al - QSAR Linear Regression on 18...	Linear Regression using 18 MeOx NPs, with selected descriptors (normalised): Standard Enthalpy...	2016-03-15T14:19:17.452+0000
walk-84-fingerprint	Walkey et al - QSAR PLS with VIP scores on 84...	PLS with VIP scores on 84 Gold NPs with 76 descriptors (Protein fingerprint; Spectral Counts) usi...	
walk-84-zp-plsvip	Walkey et al - QSAR PLS with VIP scores on 84...	PLS with VIP scores on 84 Gold NPs with 6 normalised descriptors (Zeta Potential after synthesis,...	

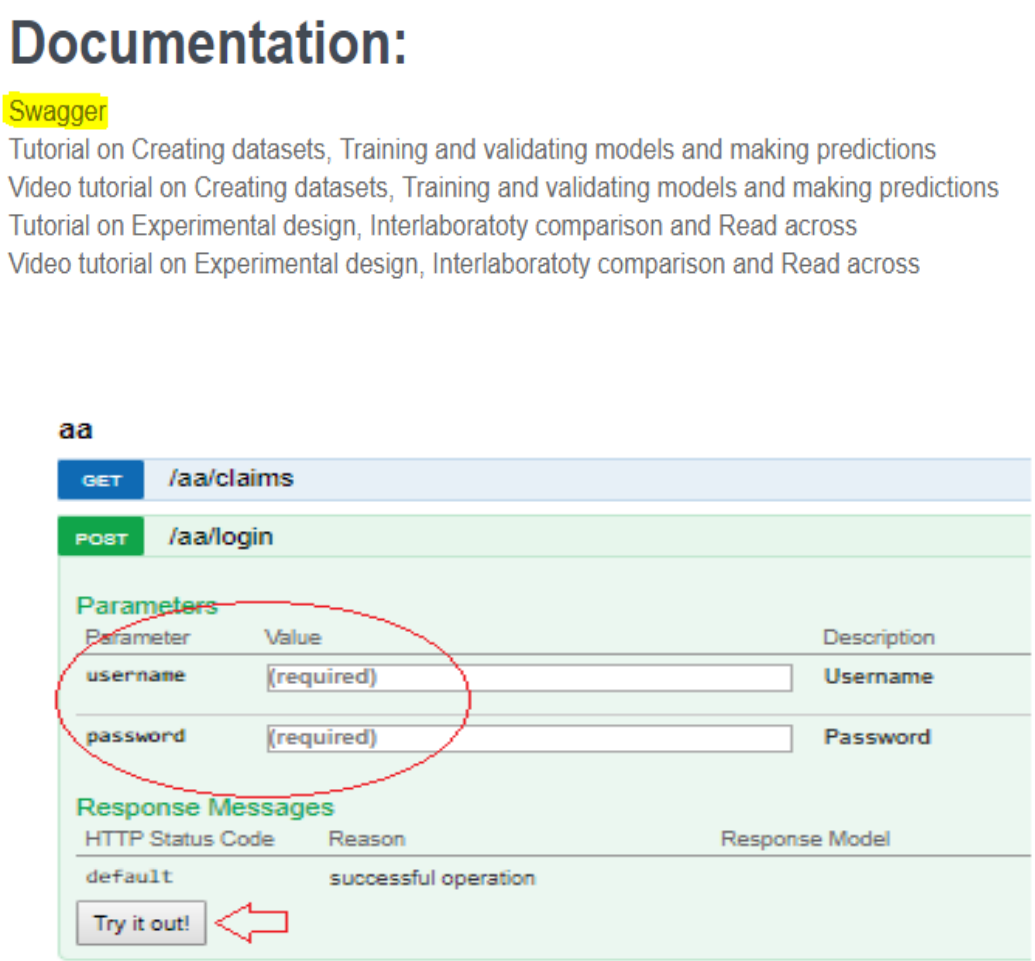
Name	Title	Description	Date
YlvjkwOYK000lGP0qZLd	Model predicting cytotoxicity of metal oxide NPS	Model developed by Mu et al in 2016	2019-05-14T06:07:51.186+0000
fU8dKE0Ej7wSFZ0hoZ16	Model predicting cytotoxicity of metal oxide NPS	Model developed by Mu et al in 2016	2019-05-06T11:44:54.168+0000
JB0T7ZUV2AxmWxdif9TN	Model predicting cytotoxicity of metal oxide NPS	Model developed by Mu et al in 2016	2019-05-06T11:43:44.750+0000
svFvHstT77WqBLoevM64	Model predicting cytotoxicity of metal oxide NPS	Model developed by Gajewicz et al in 2015	2019-05-06T11:12:30.361+0000
PBFh6o69r76dFnA4mV5e	Model predicting solubility of C60 fullerene_298K_	Model developed by Marcus et al in 2001	2019-05-06T11:03:18.754+0000
C7nlZKlt283BxgxdR71Y	Model predicting solubility of C60 fullerene_303K	Model developed by Marcus et al in 2001	2019-05-06T10:50:46.044+0000
v86ZboISbRvgbZRxuNzv	Model predicting solubility of C60 fullerene_298K	Model developed by Marcus et al in 2001	2019-05-06T10:48:45.295+0000
ZsQd750xNIq2nDRMGCBA	Model for predicting cytotoxicity of metal oxide n	Model developed by Puzyn et al in 2011	2019-05-06T09:56:57.899+0000
hhcNeubWkix3s665XXaP	Model predicting cytotoxicity for copper NPs	Model developed by Rispoli et al in 2010 (Simplex centroid design)	2019-05-06T09:34:22.687+0000
igq2gVw9Qwr7IVxjMKPa	Model predicting cytotoxicity for metal oxides	Model developed by Pathakoti et al in 2014, Photo-induced (light) case	2019-05-06T09:19:00.635+0000
w1VtlLZrZmgMMMSHQKpJ	Model predicting cytotoxicity for metal oxides	Model developed by Pathakoti et al in 2014, dark condition case	2019-05-06T09:13:06.649+0000
lGgonmAO2DEqNH94UPjz	Model predicting C60 solubility in org solv 2007	Model developed by Toporov et al in 2007	2019-05-06T06:22:48.266+0000
sCoqY3D3xCpSuyS6RdoQ	Model for predicting C60 solubility in organic sol	Model developed by Toporov et al in 2008	2019-05-06T05:48:04.408+0000
m5JcUhFafaVlUXoyZny9	Model predicting pEC50 in metal oxides, MLR, ...	Model developed by Kar et al in 2014, Strepwise MLR, mod_1	2019-05-05T23:22:22.927+0000
MHhXI2Ae7HSOE51la3pG	Model predicting pEC50 in metal oxides, MLR, ...	Model developed by Kar et al in 2014, Strepwise MLR, mod_1	2019-05-05T23:16:43.280+0000
5YXM6OzrnvjaOtGV5xHh	Model for predicting cytotoxicity of metal oxide n	Model developed by Pan et al in 2016, mod 2	2019-05-05T23:03:43.099+0000
FYtdbuUlB3vD0yQ1GaXl	Model for predicting cytotoxicity of metal oxide n	Model developed by Pan et al in 2016, mod 1	2019-05-05T22:56:25.860+0000
qQXPTMI3V3wJ5HyisMkR	Predicting Solubility of C60 fullerene	Model developed by Marcus et al in 2001 (303 K)	2019-05-05T22:20:41.893+0000

Εικόνα 19: Αποθετήριο μοντέλων

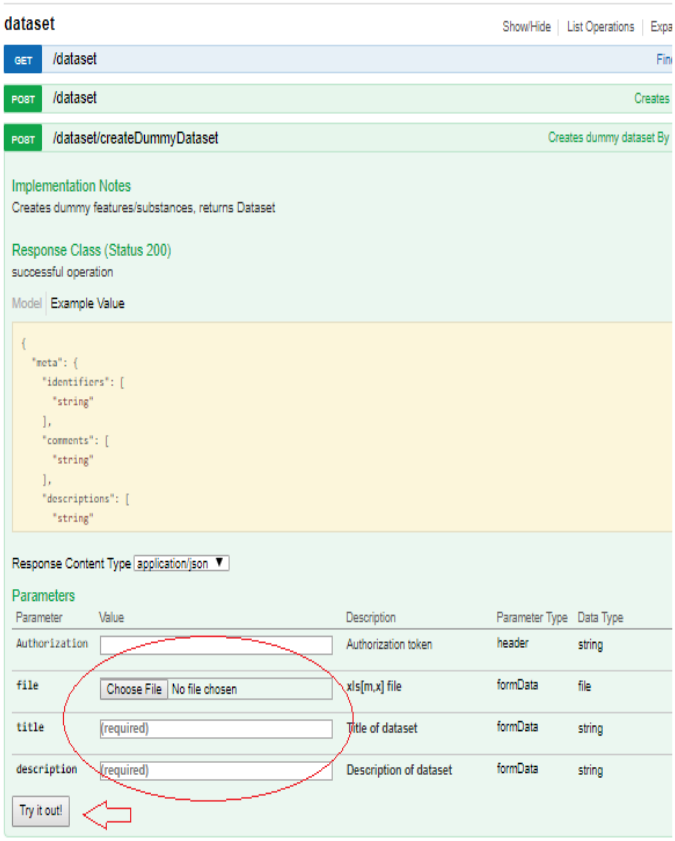
4.2.2. Μεταφόρτωση Συνόλου Δεδομένων (Upload dataset)

Αν ο χρήστης θέλει να μεταφορτώσει δεδομένα (αρχείο μορφής .csv) ώστε να τα αποθηκεύσει στο ιδιωτικό του αποθετήριο και να τα έχει διαθέσιμα τότε θα πρέπει αφού μεταβεί στο Documentation στο κάτω μέρος της Αρχικής Οθόνης (βλ. Εικόνα 17) και από εκεί στο Swagger να ακολουθήσει την παρακάτω διαδικασία (βλ Εικόνες 20-):

Documentantation → Swagger → aa → POST/aa/login → Username/ Password → Try it out → dataset → POST /dataset/createDummyDataset → Επιλογή Αρχείου, Συμπλήρωση τίτλου & περιγραφής → Try it out



Εικόνα 20: Σύνδεση



Εικόνα 21: Μεταφόρτωση αρχείου δεδομένων

Μετά την ολοκλήρωση το dataset εμφανίζεται στον προσωπικό χώρο του χρήστη (βλ. Εικόνα 19)

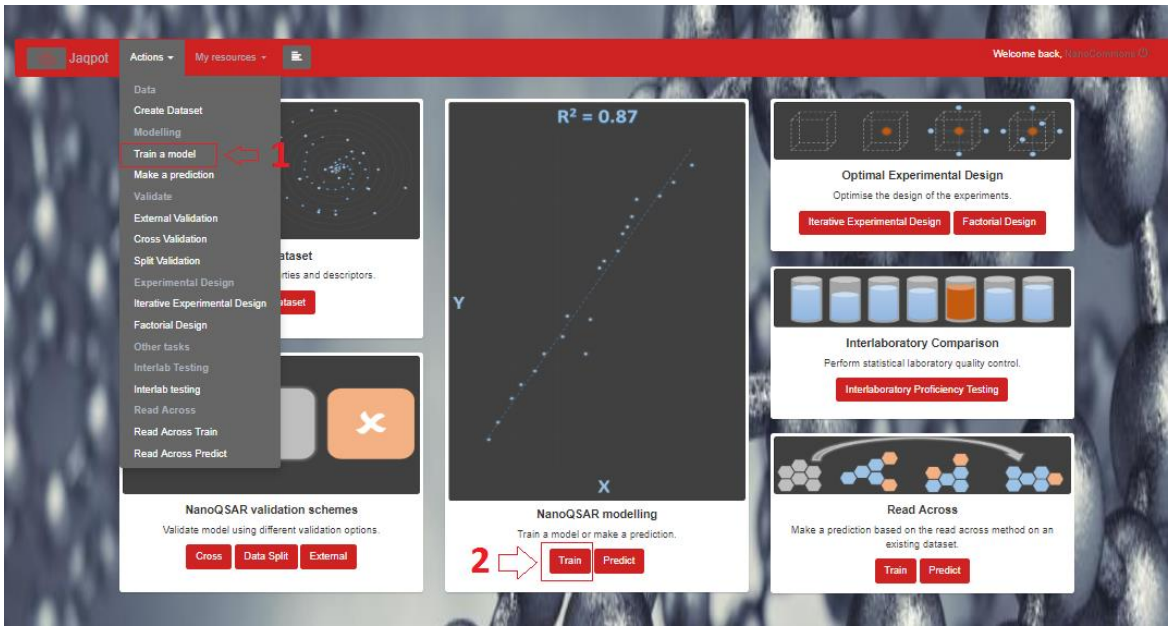
4.2.3. Δημιουργία Μοντέλου

Για την καλύτερη κατανόηση του τρόπου δημιουργίας ενός μοντέλου στο Jaqpott Quattro θα γίνει παρουσίαση ενός πραγματικού case με την υλοποίηση ενός μοντέλου της βιβλιογραφίας που προβλέπει τη διαλυτότητα C60 σε οργανικούς διαλύτες (Gharagheizi, F., & Alamdari, R. F. (2008) "Predicting C60 solubility in organic solvents by means of a molecular-based model).

Η διαδικασία δημιουργίας ενός μοντέλου μπορεί να ξεκινήσει με δύο τρόπους, όπως φαίνεται και στην Εικόνα 22:

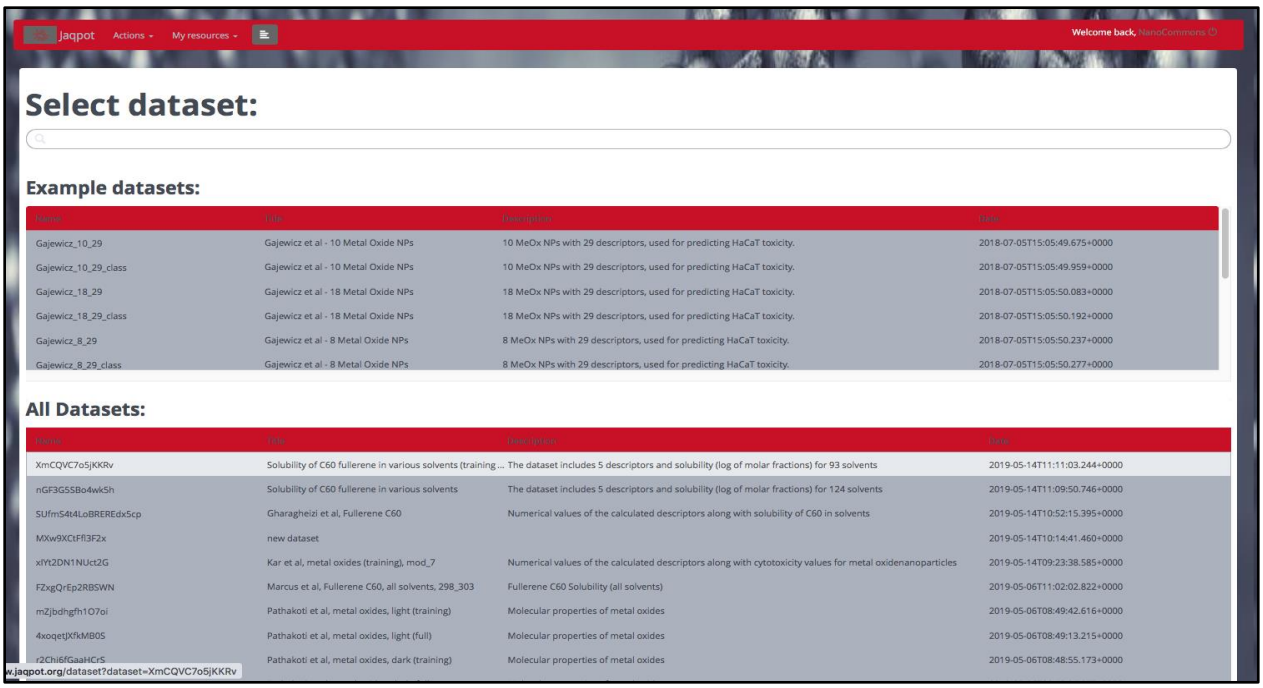
1. επιλέγοντας "Actions" στην συνέχεια "Train a model"
2. Επιλέγοντας "Train" στην ενότητα "NanoQSAR modelling" της κύριας οθόνης Jaqpot.

Στις επόμενες οθόνες ο χρήστης καλείται να επιλέξει το σύνολο δεδομένων (Εικόνα 23) και τον αλγόριθμο (Εικόνα 24) που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου. Ένα από τα βασικά κριτήρια στην επιλογή του αλγορίθμου είναι αν το πρόβλημα είναι ταξινόμησης ή παλινδρόμησης.



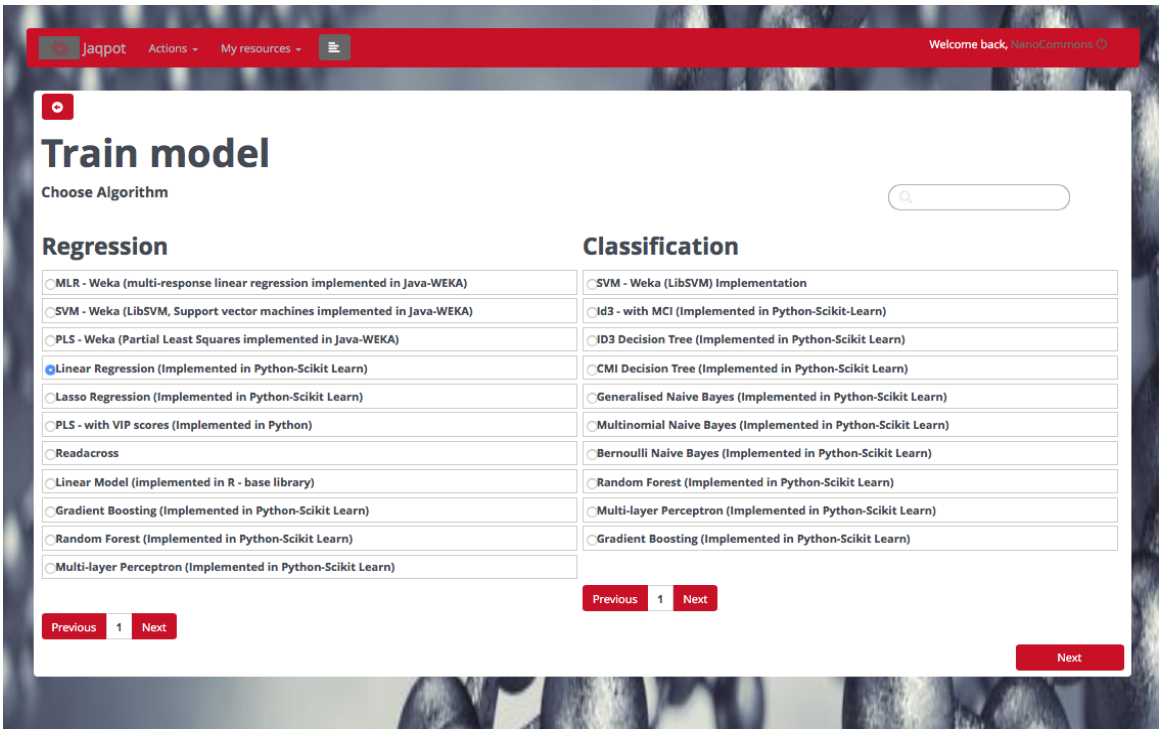
Εικόνα 22: Έναρξη μοντελοποίησης

Στο επόμενο βήμα ο χρήστης καλείται να επιλέξει το σύνολο δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου (training dataset). Στην συγκεκριμένη περίπτωση το σύνολο δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου αποτελείται από 93 διαλύτες και είναιδιαθέσιμο στην ηλεκτρονική διεύθυνση http://www.jaqpot.org/data_detail?name=XmCQVC7o5jKKRv οπότε το επιλέγουμε:



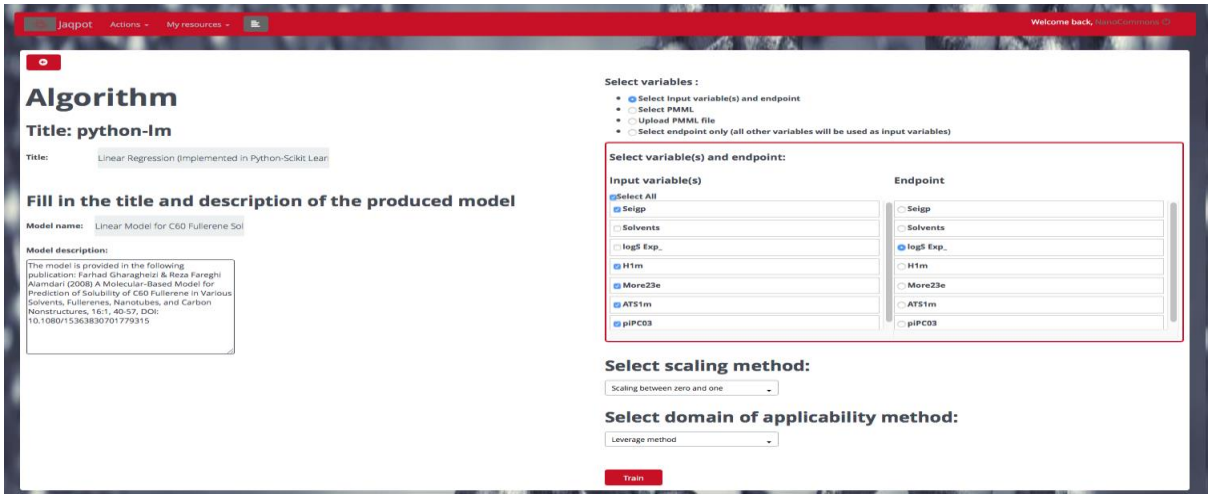
Εικόνα 23: Επιλογή συνόλου δεδομένων

Η επόμενη οθόνη παροτρύνει τον χρήστη να επιλέξει από την βιβλιοθήκη των διαθέσιμων αλγορίθμων εκείνον που θα χρησιμοποιηθεί για την υλοποίηση του μοντέλου. Στην συγκεκριμένη περίπτωση επιλέγεται ο αλγόριθμοςLinear Regression.



Εικόνα 24: Επιλογή αλγορίθμου

Αμέσως μετά την επιλογή του αλγορίθμου, ο χρήστης μεταφέρεται στην επόμενη οθόνη (Εικόνα 25) όπου και καλείται να επιλέξει τις μεταβλητές εισόδου και την μεταβλητή απόκρισης αλλά και να συμπληρώσει επιπλέον πεδία σχετικά με το μοντέλο.

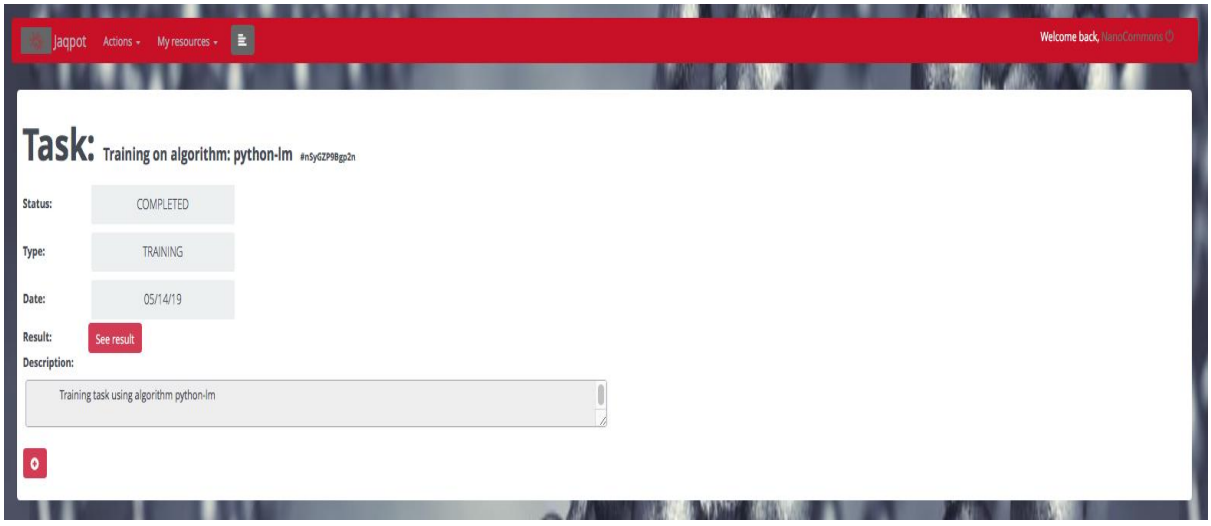


Εικόνα 25: Παράμετροι αλγορίθμου, λεπτομέρειες μοντέλου και επιλογή μεταβλητών

Επεξήγηση πεδίων:

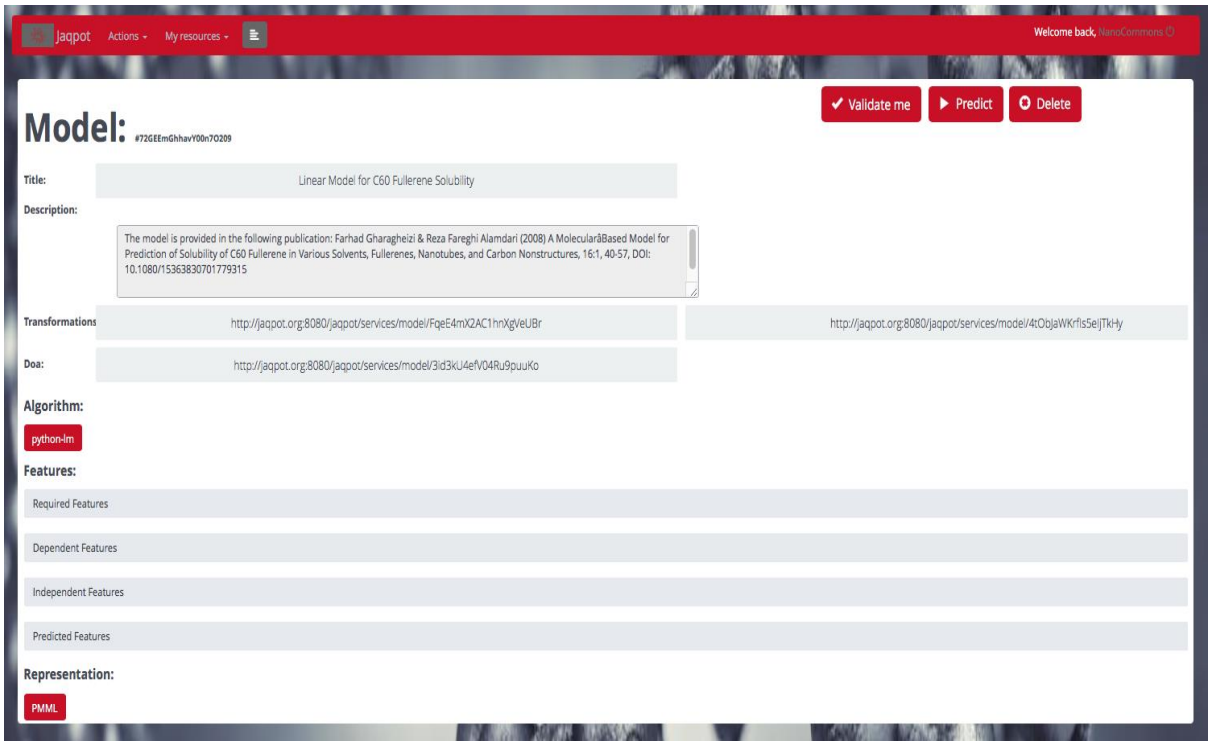
- **Title of the algorithm:** Ο τίτλος του αλγορίθμου που χρησιμοποιείται για την υλοποίηση του μοντέλου, συμπληρώνεται αυτόματα από το σύστημα σύμφωνα με την επιλογή του προηγούμενου βήματος.
- **Model name**: Ο τίτλος του μοντέλου.
- **Model description**: Η περιγραφή του μοντέλου.
- **Select variables**: Επιλογή των μεταβλητών που θα χρησιμοποιηθούν για το μοντέλο. Ορίζονται τόσο οι ανεξάρτητες όσο και οι εξαρτημένες.
- **Select scaling method**: Κλιμάκωση. Στάδιο προεπεξεργασίας των μεταβλητών που στόχο έχει την αποφυγή μεγάλων αποκλίσεων. Παρέχεται η δυνατότητα επιλογής ανάμεσα σε κλιμάκωση από 0 εως 1 και κανονικοποίηση.
- **Select domain of applicability method**: Τομέας εφαρμογής. DoA παρέχουν ένα μέτρο της ικανότητας του μοντέλου να παρέχει αξιόπιστες προβλέψεις για κάθε προβλεπόμενο σημείο.

Μετά την συμπλήρωση των παραπάνω πεδίων και την επιλογή «Train» εκκινείται η διαδικασία δημιουργίας του μοντέλου και ο χρήστης μεταφέρεται σε μια ενδιάμεση οθόνη εως ότου η διαδικασία αυτή ολοκληρωθεί όπως φαίνεται και στην Εικόνα 26.



Εικόνα 26: Ενδιάμεση οθόνη

Όταν ολοκληρωθεί η παραπάνω διεργασία, δηλαδή όταν το μοντέλο έχει δημιουργηθεί, το πλήκτρο «See result» είναι πλέον ενεργοποιημένο και επιλέγοντάς το ο χρήστης μεταφέρεται στην σελίδα του μοντέλου όπου του δίνοται οι επιλογές επικύρωσης του μοντέλου, χρήσης του μοντέλου για πρόβλεψη και διαγραφής του μοντέλου όπως φαίνεται και στην Εικόνα 27.



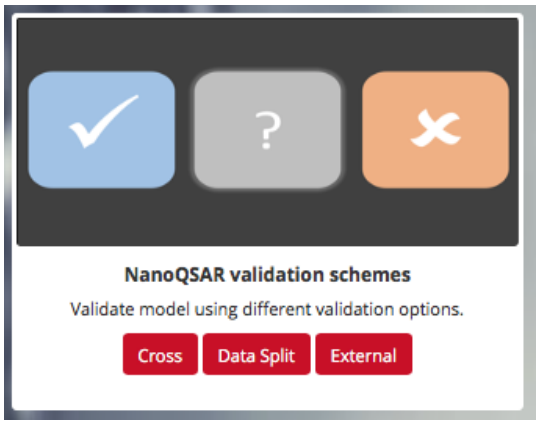
Εικόνα 27: Σελίδα του μοντέλου που δημιουργήθηκε

Αναφέρθηκε νωρίτερα πως τόσο τα σύνολα δεδομένων όσο και τα μοντέλα που δημιουργούνται, προστίθενται στο αποθετήριο του εργαλείου ώστε να είναι διαθέσιμα σε όποιον θέλει να τα χρησιμοποιήσει. Στην συγκεκριμένη περίπτωση το μονέλο που δημιουργήθηκε είναι διαθέσιμο στην ηλεκτρονική διεύθυνση: http://jaqpot.org/m_detail?name=sCoqY3D3xCpSuyS6RdoQ.

4.2.4. Επικύρωση Μοντέλου (Validation)

Όπως έχει ήδη αναφερθεί το Jaqpot Quattro διαθέτει τρεις επιλογές για την επικύρωση των παραγόμενων μοντέλων μέσω του συστήματος δυνατοτήτων επικύρωσης όπως φαίνεται και στην Εικόνα 28. Πρόκειται για τις κάτωθι:

1. **External validation**: Πραγματοποιείται με την χρήση ενός εξωτερικού συνόλου δεδομένων.
2. **Cross validatation**: Διασταυρούμενη επικύρωση που έχει περιγραφεί αναλυτικά σε προηγούμενο κεφάλαιο.
3. **Split validation**: Πραγματοποιείται με τον διαχωρισμό του συνόλου δεδομένων σε δύο επιμέρους σύνολα (test και data) σύμφωνα με έναν λόγο διαίρεσης.

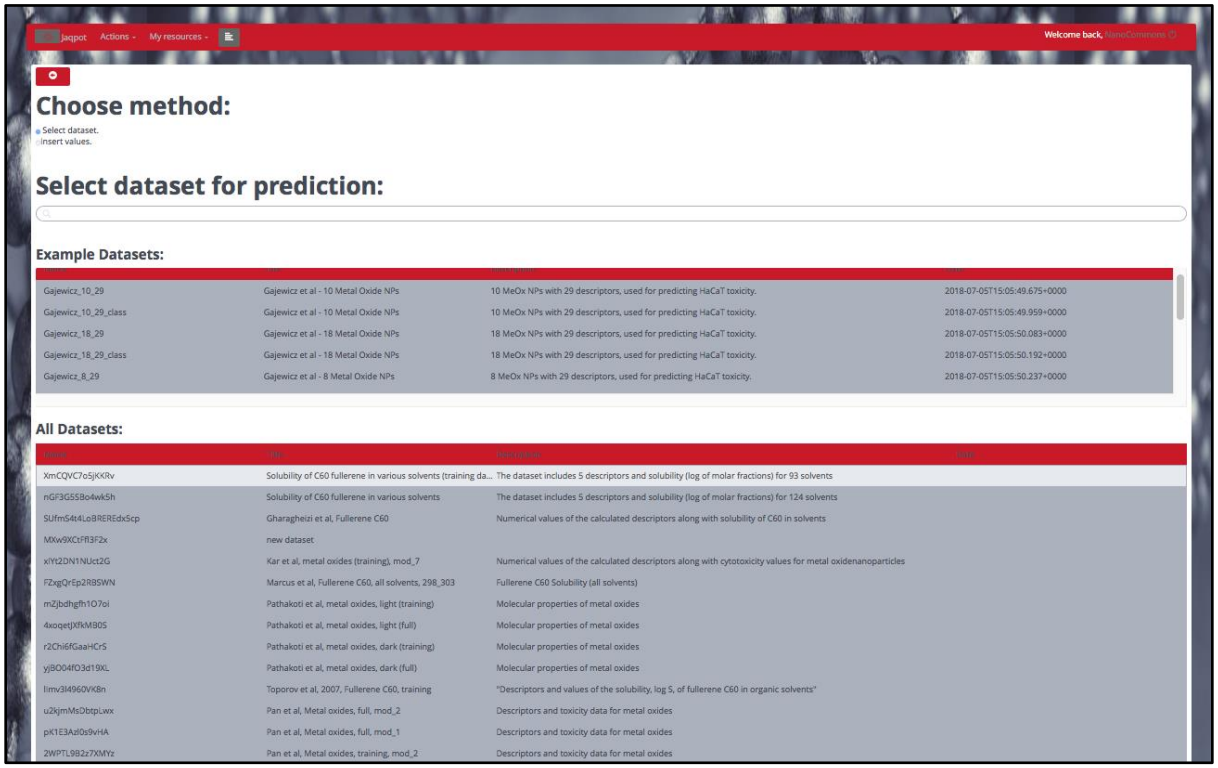


Εικόνα 28: Επιλογές Επικύρωσης Μοντέλου

4.2.4.1. Εξωτερική Επικύρωση (External Validation)

Όταν πλέον το μοντέλο έχει δημιουργηθεί έχουμε πλέον την δυνατότητα να επικυρώσουμε/αξιολογήσουμε την εγκυρότητά του. Αυτό μπορεί να πραγματοποιηθεί πολύ εύκολα επιλέγοντας το πλήκτρο «Validate me» που φαίνεται στην Εικόνα 27. Τότε μεταφερόμαστε στην οθόνη της Εικόνας 29 όπου καλούμαστε να επιλέξουμε τον τρόπο εισαγωγής των δεδομένων που θα χρησιμοποιηθούν για την επικύρωση του μοντέλου. Το εργαλείο όπως φαίνεται παρέχει δύο επιλογές.

1. **Select dataset:** Όπου επιλέγεται ένα από τα διαθέσιμα σύνολα δεδομένων
2. **Insert values:** Όπου εμφανίζεται ένα ενσωματωμένο υπολογιστικό φύλλο στο οποίο ο χρήστης μπορεί να πληκτρολογήσει τις τιμές του συνόλου δεδομένων ή να αντιγράψει και να επικολλήσει τιμές από ένα υπολογιστικό φύλλο στις αντίστοιχες στήλες (Εικόνα 29).

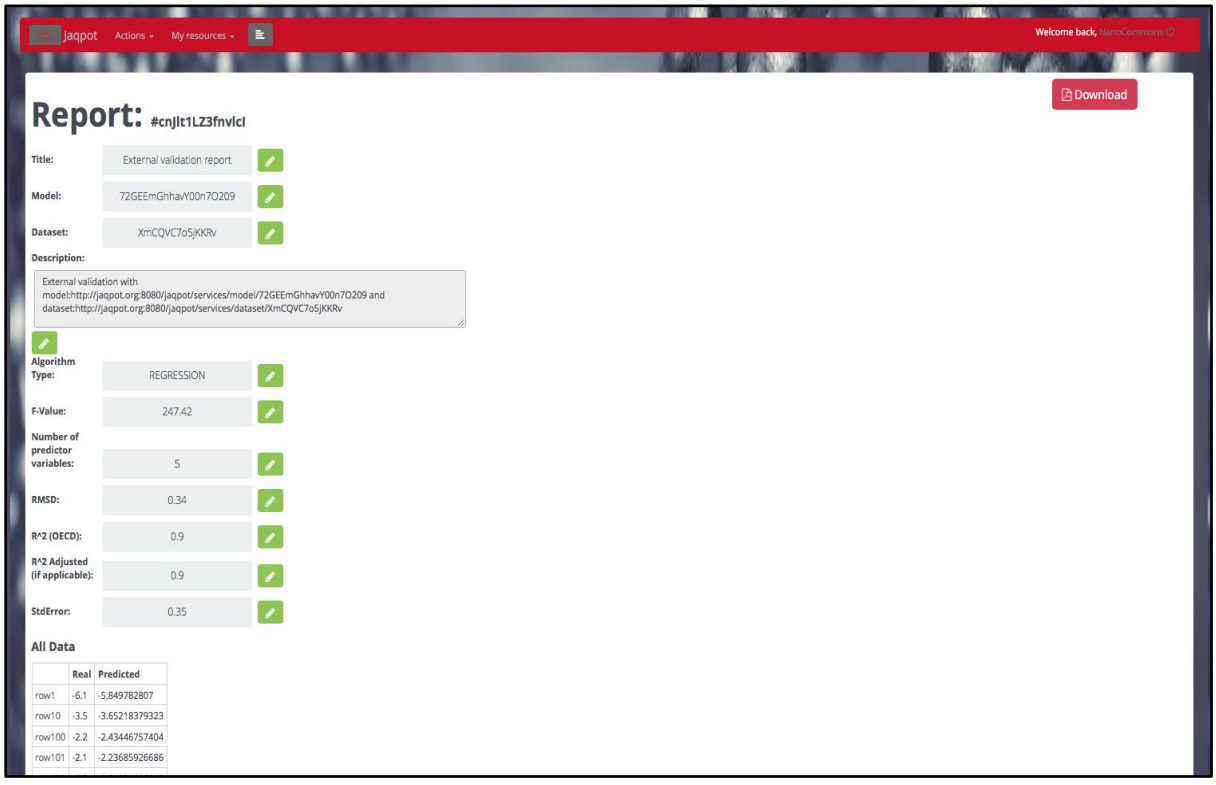


Εικόνα 29: Επιλογή δεδομένων για την επικύρωση του μοντέλου

Στην περίπτωση του δικού μας μοντέλου επιλέξαμε την δεύτερη επιλογή και με την λειτουργία της αντιγραφής – επικόλλησης εισάγαμε τα δεδομένα ελέγχου (test dataset).

Εικόνα 30: Εισαγωγή τιμών για για την επικύρωση του μοντέλου

Κλικάροντας το πλήκτρο «Validate» ξεκινά η διαδικασία επικύρωσης και μόλις αυτή ολοκληρωθεί το σύστημα παρέχει στον χρήστη μια αναφορά επικύρωσης (Validation Report) που περιλαμβάνει τις σχετικές μετρήσεις επικύρωσης, έναν πίνακα σύγκρισης μεταξύ των πραγματικών τιμών και των τιμών που προβλέφθηκαν με την χρήση του μοντέλου καθώς και μια γραφική παράσταση όπως φαίνεται και στις ακόλουθες εικόνες. Ενώ η εν λόγω αναφορά δύναται να αποθηκευθεί και σε αρχείο μορφής pdf (Εικόνα 33).



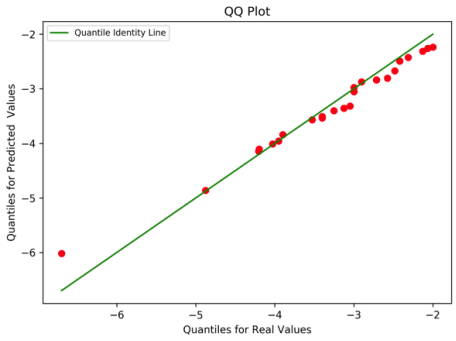
Εικόνα 31: Αναφορά Επικύρωσης_1 (External Validation)

All Data

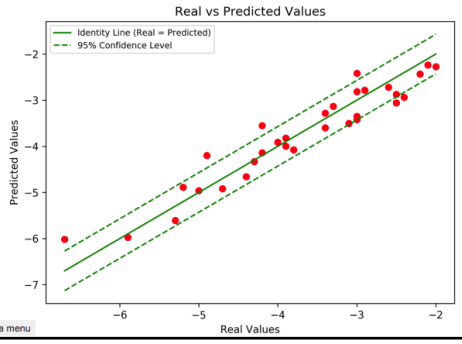
	Real	Predicted
compound1	-4	-3.91355789467
compound10	-2.6	-2.72064673994
compound11	-4.9	-4.19945022239
compound12	-5	-4.96443038285
compound13	-2.1	-2.23685926686
compound14	-3	-3.34976177746
compound15	-2	-2.27234329025
compound16	-3.9	-3.82107851451
compound17	-5.2	-4.88772442914
compound18	-4.2	-3.5504783112
compound19	-6.7	-6.01794584696
compound2	-5.3	-5.61018322885
compound20	-4.3	-4.33212597512
compound21	-2.5	-3.05827270109
compound22	-3.3	-3.13183407097
compound23	-2.4	-2.93903145294
compound24	-3.8	-4.07882245078
compound25	-3	-2.81223578464
compound26	-4.4	-4.65814301486
compound27	-4.7	-4.91962777437
compound28	-2.5	-2.87194521299
compound29	-3.9	-3.99928281415
compound3	-3.4	-3.28038430934
compound30	-2.2	-2.43446757404
compound31	-3	-3.41838538257
compound4	-5.9	-5.97659581681
compound5	-4.2	-4.14080958734
compound6	-3.1	-3.5059080608
compound7	-2.9	-2.78268289953
compound8	-3.4	-3.59838096199
compound9	-3	-2.41555106252

Εικόνα 32: Αναφορά Επικύρωσης_2 (External Validation)

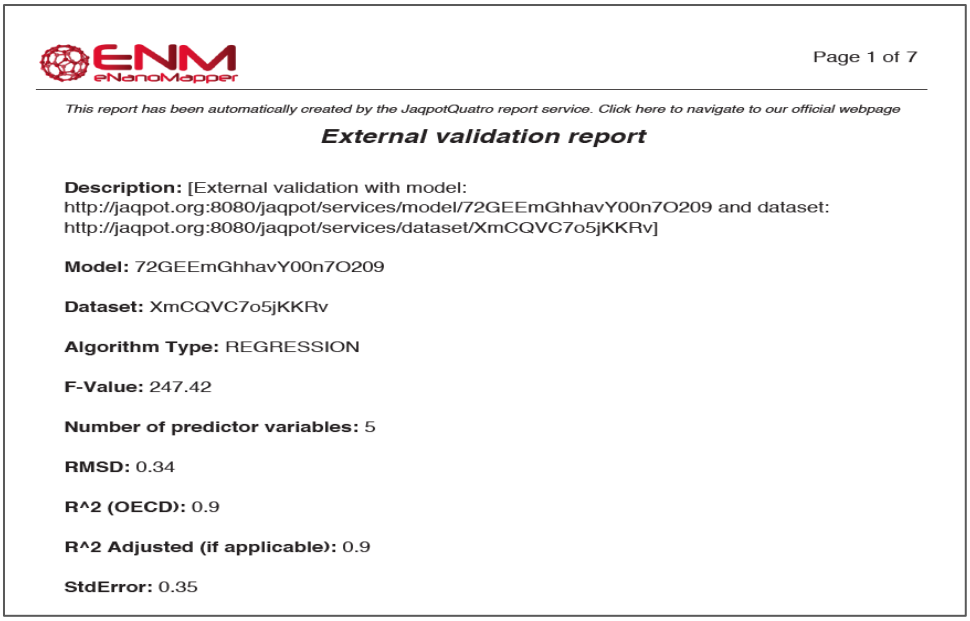
QQ Plot



Real Vs Predicted



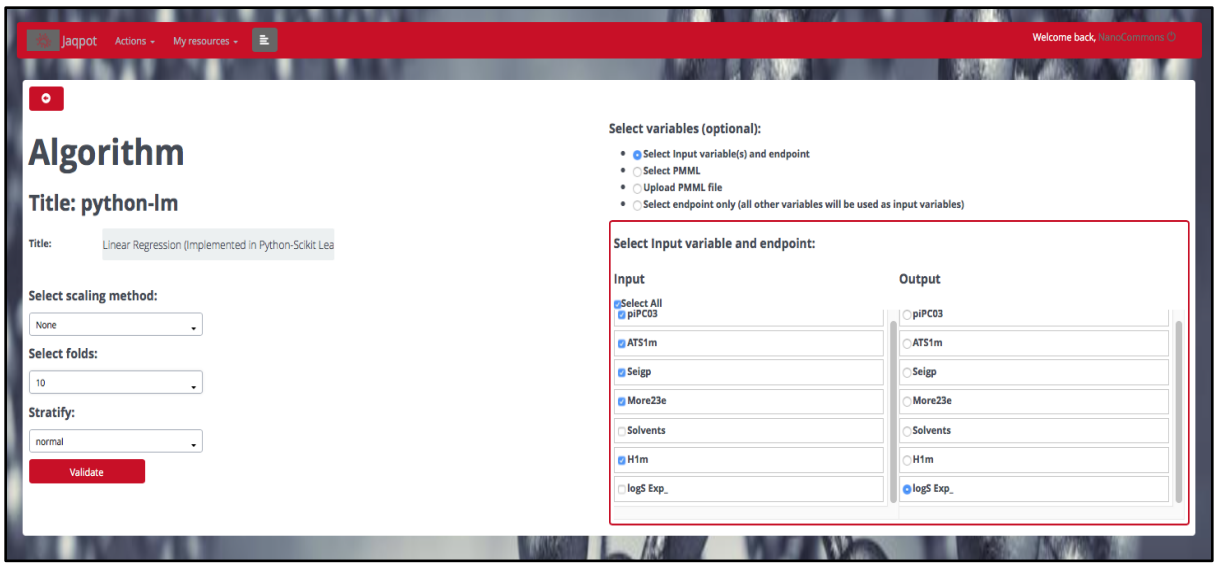
Εικόνα 33: Αναφορά Επικύρωσης_3 (External Validation)



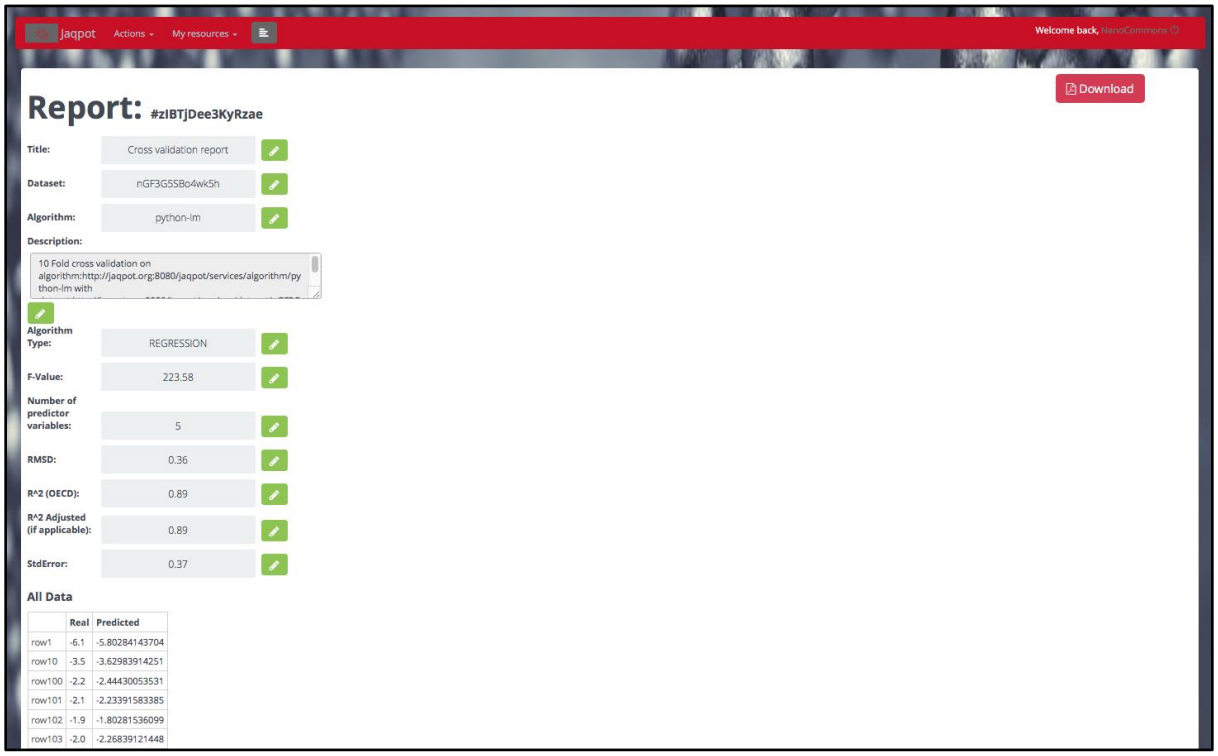
Εικόνα 34: Αναφορά Επικύρωσης σε μορφή pdf (External Validation)

4.2.4.2. Διασταυρούμενη Επικύρωση (Cross Validation)

Η διαδικασία της Διασταυρούμενης Επικύρωσης, διαφέρει σημαντικά από αυτή της εξωτερικής και προσομοιάζει περισσότερο την διαδικασία δημιουργίας ενός μοντέλου. Αφού ο χρήστης επιλέξει το σύνολο δεδομένων, καλείται να ορίσει τα δεδομένα του αλγορίθμου που επιθυμεί να επικυρώσει όπως φαίνεται στην Εικόνα 35. Η διαφορά που παρουσιάζεται σε σχέση με την εξαρχής δημιουργία του μοντέλου είναι πως απαιτείται η συμπλήρωση επιπλέον παραμέτρων όπως ο ορισμός του αριθμού των πτυχών που θα χρησιμοποιηθούν στην διαδικασία διασταυρούμενης επικύρωσης καθώς και η επιλογή του είδους διαστρωμάτωσης, ανάμεσα σε τυχαία, κανονική ή καθόλου. Στην συνέχεια ο χρήστης κλικάροντας το πλήκτρο «Validate» και λαμβάνει απ' ευθείας το validation report (Εικόνα 36), το οποίο παρουσιάζει πολλές ομοιότητες με αυτό που παράγεται κατά την διαδικασία της εξωτερικής επικύρωσης.



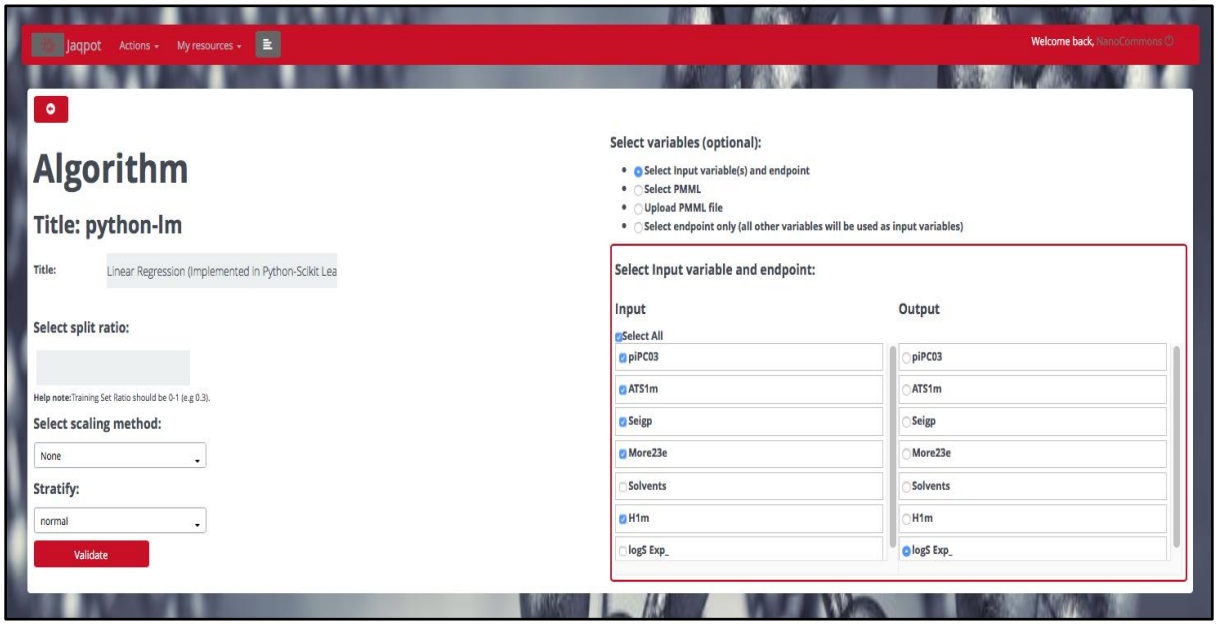
Εικόνα 35: Παράμετροι αλγορίθμου, λεπτομέρειες μοντέλου και επιλογή μεταβλητών κατά τη διάρκεια διασταυρούμενης επικύρωσης



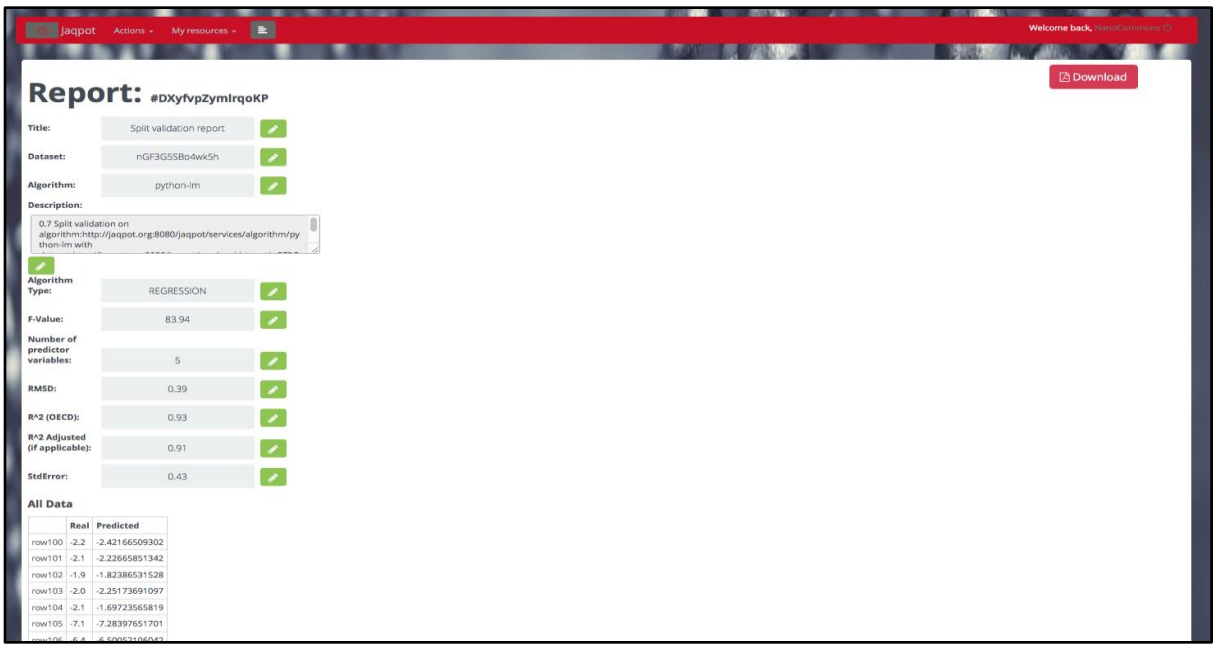
Εικόνα 36: Αναφορά Επικύρωσης (Cross Validation)

4.2.4.3. Επικύρωση με διαχωρισμό δεδομένων (Split Validation)

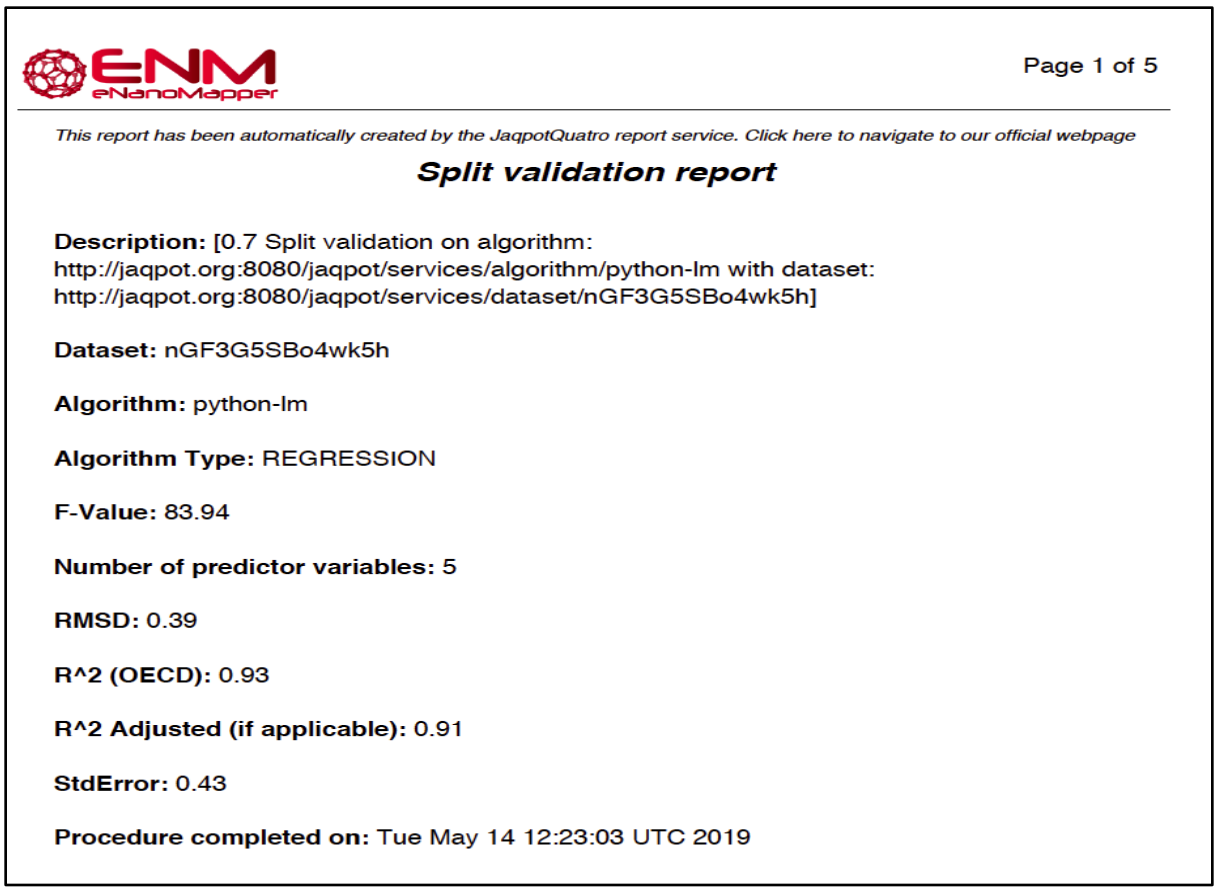
Η επιλογή της δυνατότητας επικύρωσης με διαχωρισμό δεδομένων ανακατευθύνει τον χρήστη σε μια οθόνη όπου καλείται να ορίσει τα δεδομένα του αλγορίθμου που επιθυμεί να επικυρώσει, την αναλογία διαχωρισμού και τις επιλογές σχετικά με την διαστρωμάτωση όπως φαίνεται στην Εικόνα 37. Στην συνέχεια ο χρήστης κλικάροντας το πλήκτρο «Validate» και λαμβάνει απ' ευθείας το validation report (Εικόνα 38), το οποίο αν θέλει μπορεί κλικάροντας «Download» πάνω δεξιά να το αποθηκεύσει σε μορφή pdf (Εικόνα 39).



Εικόνα 37: Παράμετροι αλγορίθμου, λεπτομέρειες μοντέλου και επιλογή μεταβλητών κατά τη διάρκεια της επικύρωσης με διαχωρισμό δεδομένων



Εικόνα 38: Αναφορά Επικύρωσης (Split Validation)



Page 1 of 5

This report has been automatically created by the JaqpotQuatro report service. Click here to navigate to our official webpage

Split validation report

Description: [0.7 Split validation on algorithm:
http://jaqpot.org:8080/jaqpot/services/algorithm/python-lm with dataset:
http://jaqpot.org:8080/jaqpot/services/dataset/nGF3G5SBo4wk5h]

Dataset: nGF3G5SBo4wk5h

Algorithm: python-lm

Algorithm Type: REGRESSION

F-Value: 83.94

Number of predictor variables: 5

RMSD: 0.39

R^2 (OECD): 0.93

R^2 Adjusted (if applicable): 0.91

StdError: 0.43

Procedure completed on: Tue May 14 12:23:03 UTC 2019

Εικόνα 39: Αναφορά Επικύρωσης σε μορφή pdf (Split Validation)

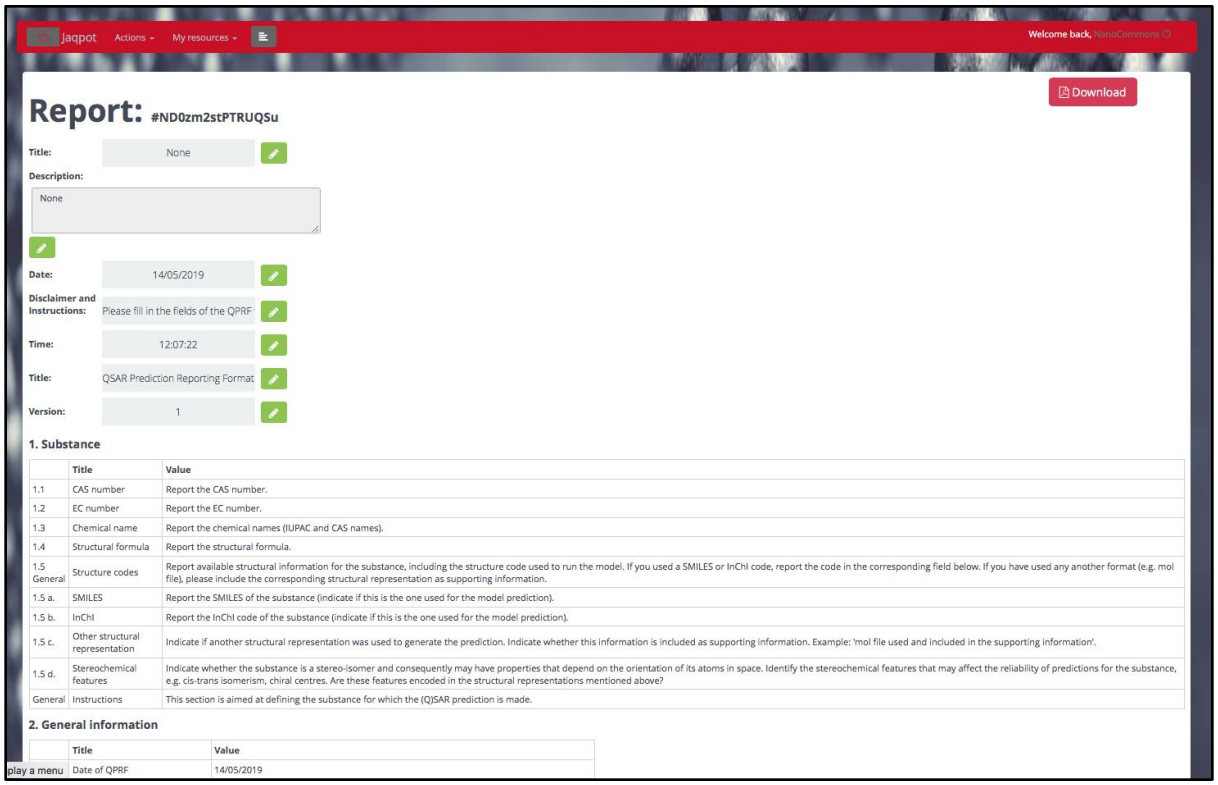
4.2.5. Πρόβλεψη τιμών μεταβλητής απόκρισης (Prediction of endpoint)

Έχοντας ολοκληρώσει την διαδικασία επικύρωσης, ο χρήστης είναι πλέον σε θέση να χρησιμοποιήσει το μοντέλο που δημιούργησε για την πραγματοποίηση προβλέψεων. Κάνοντας κλικ στο πλήκτρο «Predict» επιλέγει το σύνολο δεδομένων, όπως ακριβώς και στην διαδικασία Επικύρωσης και στη συνέχεια αφού κλικάρει ξανά το πλήκτρο «Predict» μεταφέρεται στην οθόνη που εμφανίζεται στην Εικόνα 40 η σελίδα στο Σχήμα 21, όπου παρουσιάζονται οι τιμές που προβλέφθηκαν με την χρήση του μοντέλου, οιυπολογισθέντες τιμές DoA καθώς και πλήκτρα για την αυτόματη δημιουργία αναφορών QPRF.

Εικόνα 40: Προβλεπόμενες τιμές, τιμές DoA και πλήκτρα για δημιουργία αναφοράς QPRF

Σχετικά με τις τιμές DoA (Domain of Applicability), προσδιορίζουν το κατά πόσο το στοιχείο στο οποίο αναφέρεται η πρόβλεψη είναι εντος του πεδίου εφαρμογής του μοντέλου άρα πρέπει να λαμβάνεται υπόψη. Προβλέψεις με τιμές κοντά στο 0 δηλώνουν στοιχείο εκτός του πεδίου εφαρμογής, άρα μη αξιόπιστη πρόβλεψη, ενώ όσο αυξάνεται η απόσταση της τιμής DoA από 0, τόσο αυξάνεται η αξιοπιστία της πρόβλεψης αφού πλέον αναφερόμαστε σε στοιχείο εντός του πεδίου εφαρμογής του μοντέλου.

Σχετικά με την αναφορά QPRF, δυνατότητα παραγωγής της οποίας προσφέρει το Jaqpot όπως φαίνεται στην Εικόνα 41, επισημαίνεται πως συμμορφώνεται πλήρως με τις κατευθυντήριες οδηγίες του OECD καθώς περιέχει όλα τα υποχρεωτικά από τις οδηγίες πεδία και συγκεκριμένα πληροφορίες για τις υπό μελέτη ουσίες, τον δημιουργό του μοντέλου, το μοντέλο (μεταβλητές εισόδου, μεταβλητές απόκρισης, αλγόριθμος, DoA, κλπ) και την επάρκεια του.



Εικόνα 41: Παραγόμενη από το Jaqpot αναφορά QPRF

4.3. Jaqpot v5

Το Jaqpot v5 αποτελεί εξέλιξη του Jaqpot Quattro και έχοντας πιο σύγχρονο και φιλικό στον χρήστη γραφικό περιβάλλον προσφέρει βελτιωμένη εμπειρία αυξημένες επιλογές ανταλλαγής δεδομένων και μοντέλων καθιστώντας την συνεργασία μεταξύ διαφόρων ομάδων εξαιρετικά εύκολη. Σκοπός της δημιουργίας του ήταν να μπορσει να ανταπεξέλθει στον αυξημένο αριθμό χρηστών, τον αυξημένο όγκο δεδομένων αλλα και να προσφέρει στην επιστημονική κοινότητα περσσότερες επιλογές όσον αφορά τη χρήση αλγορίθμων. Στην παρούσα φάση το Jaqpot έχει τη δυνατότητα ενσωμάτωσης της πιο δημοφιλούς και ολοκληρωμένης βιβλιοθήκης ανοιχτού κώδικα που χρησιμοποιείται στην μηχανική μάθηση, της Python Scikit-learn ενώ προγραμματίζεται και η ενσωμάτωση αλγορίθμων και από άλλες βιβλιοθήκες (π.χ. γλώσσας R).

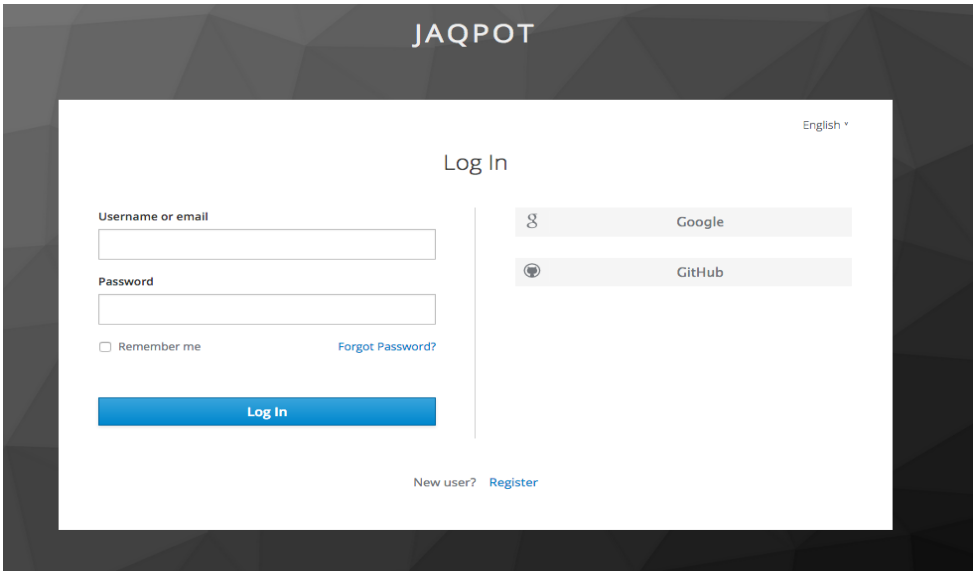
Για τη χρήση του εργαλείου Jaqpot v5 ο χρήστης αρκεί να εγκαταστήσει την βιβλιοθήκη jaqpotpy, σύμφωνα με τις αναλυτικές οδηγίες που παρέχονται στην ηλεκτρονική διεύθυνση https://jaqpotpy.readthedocs.io/en/latest, η οποία δημιουργήθηκε από την Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών ΕΜΠ και η οποία επιτρέπει την ενσωμάτωση της βιβλιοθήκης αλγορίθμων Python Scikit-learn παρέχοντας στον χρήστη την δυνατότητα να δημιουργήσει ένα μοντέλο σε γλώσσα Python και στην συνέχεια αφού το αξιολογήσει/επικυρώσει να το ανεβάσει στο περιβάλλον του Jaqpot v5 ώστε να το χρησιμοποιήσει για τις προβλέψεις του ή να το μοιραστεί με άλλους.

4.3.1. Σύνδεση/Εγγραφή & Αρχικές οθόνες

Η εφαρμογή είναι διαθέσιμη στην ηλεκτρονική διεύθυνση: https://app.jaqpot.org/ και η πρώτη εικόνα είναι η παρακάτω:

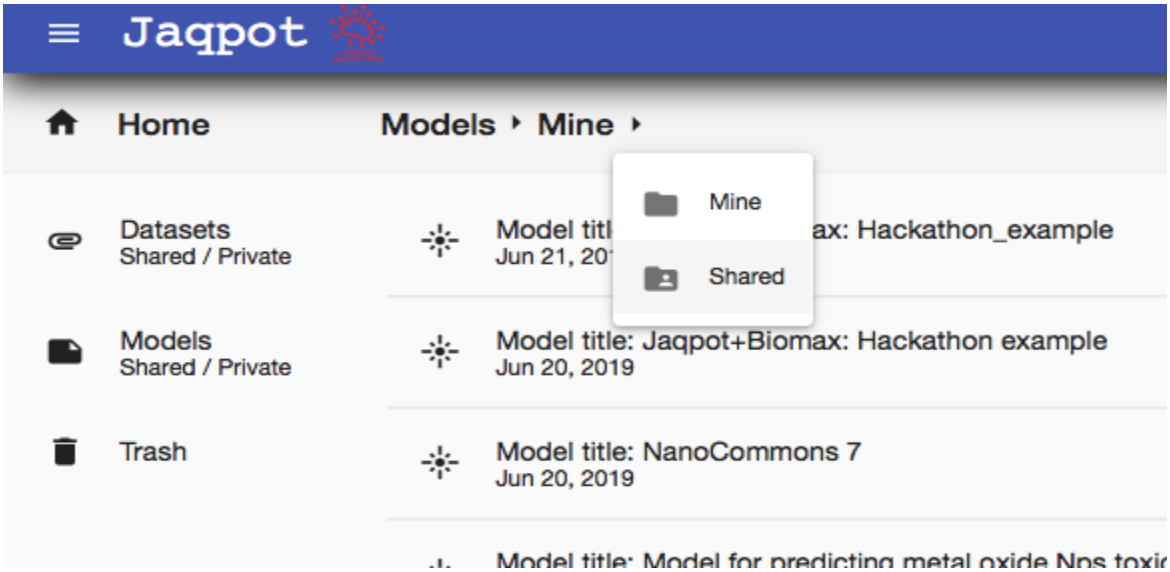
Εικόνα 42: Αρχική Οθόνη

Για την χρήση της πλατφόρμας απαιτείται η σύνδεση ή η δημιουργία λογαριασμού, οπότε κάνοντας κλικ στο εικονίδιο που βρισκεται πάνω δεξιά όπως φαίνεται στην Εικόνα 42, ο χρήστης μεταφέρεται στην επόμενη οθόνη (Εικόνα 43) όπου μπορεί να συνδεθεί αν έχει ήδη λογαριασμό ή να δημιουργήσει έναν νέο σε περίπτωση που δεν διαθέτει. Εδώ αξίζει να αναφέρουμε ότι ο χρήστης έχει την δυνατότητα να συνδεθεί χρησιμοποιώντας τους ήδη υπάρχοντες λογαριασμούς του στο Google ή στο Github δυνατότητα που δεν υπάρχει στο Jaqpot Quattro.



Εικόνα 43: Σελίδα Σύνδεσης

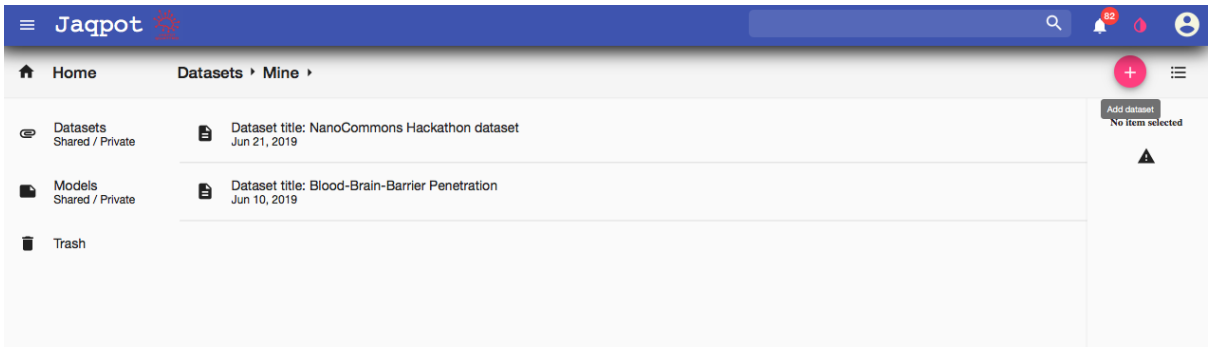
Μετά την δημιουργία του λογαριασμού ο χρήστης έχει πρόσβαση σε όλες τις λειτουργίες του εργαλείου αλλά και στο προσωπικό του αποθετήριο. Εκεί μπορεί να έχει πρόσβαση τόσο στα μοντέλα και σύνολα δεδομένων που ο ίδιος έχει δημιουργήσει, όσο και σε αυτά που είναι κοινόχρηστα στους οργανισμούς που ανήκει (Εικόνα 44)



Εικόνα 44: Αποθετήριο μοντέλων (κοινόχρηστων και προσωπικών)

4.3.2. Μεταφόρτωση Συνόλου Δεδομένων (Upload dataset)

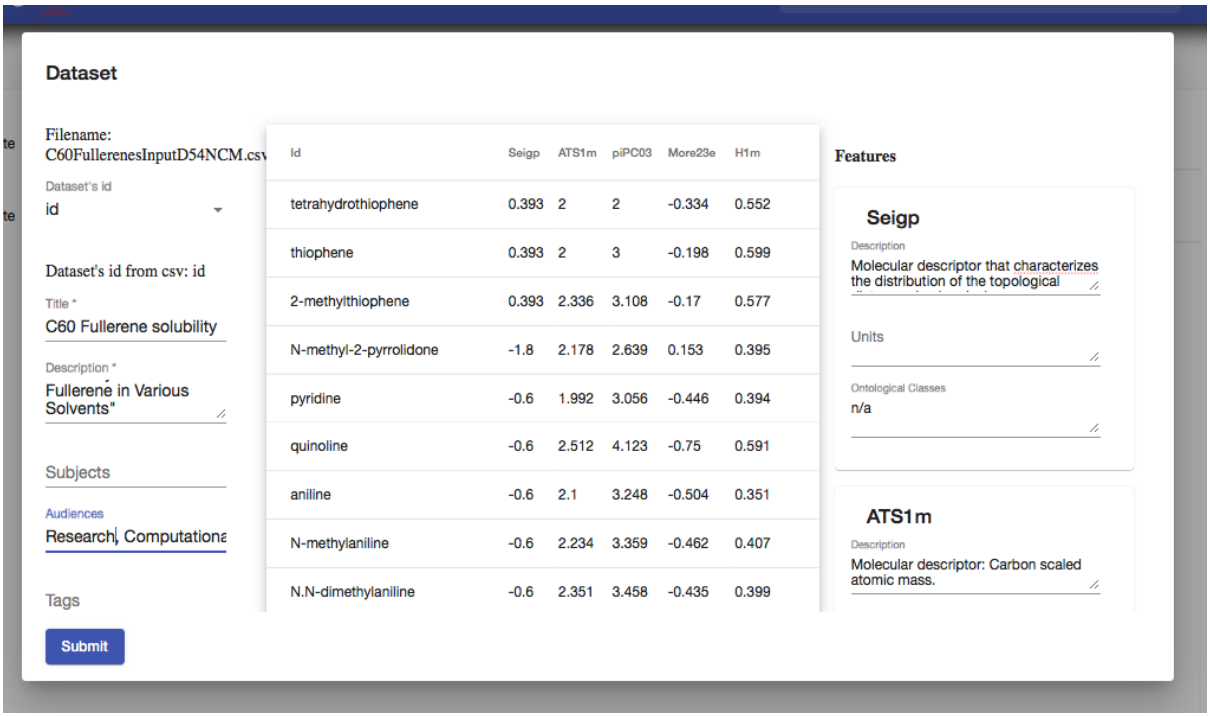
Αν ο χρήστης θέλει να μεταφορτώσει δεδομένα (αρχείο μορφής .csv) ώστε να τα αποθηκεύσει στο ιδιωτικό του αποθετήριο και να τα έχει διαθέσιμα τότε θα πρέπει να μεταβεί στην καρτέλα Datasets και να επιλέξει την επιλογή "Add dataset" όπως φαίνεται στην Εικόνα 45.



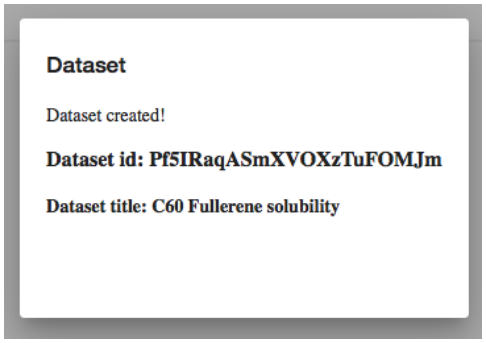
Εικόνα 45: Προσθήκη νέου Dataset

Αφού επιλεγεί το αρχείο CSV, ο χρήστης καλείται να επιλέξει το αναγνωριστικό του συνόλου δεδομένων, αυτόν το ρόλο μπορεί να παίξει μια στήλη του αρχείου CSV ή να μην υπάρχει και καθόλου (Εικόνα 46) αλλά και να συμπληρώσει κάποιες επιπλέον πληροφορίες αν και εφόσον το επιθυμεί (Εικόνα 47). Αν η παραπάνω διαδικασία ολοκληρωθεί χωρίς προβλήματα ο χρήστης ενημερώνεται για την επιτυχημένη μεταφόρτωση μέσω ένός αναδυόμενου παραθύρου που εμφανίζει τον τίτλο και το αναγνωριστικό του συνόλου δεδομένων (Εικόνα 48) και το σύνολο δεδομένων εμφανίζεται στο αποθετήριο του χρήστη (Εικόνα 49).

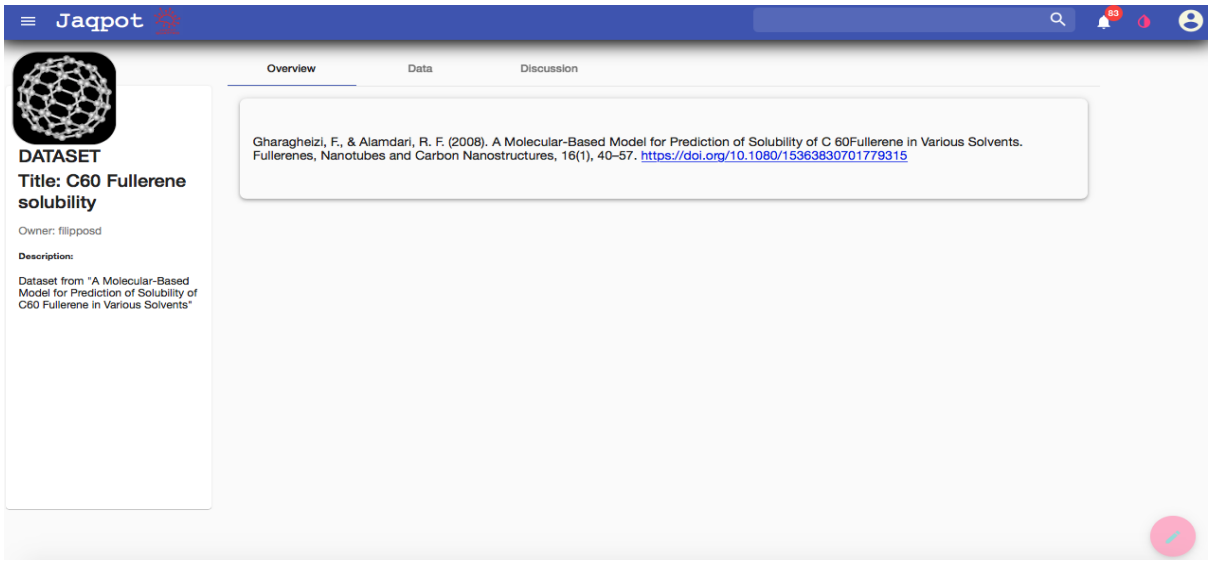
Εικόνα 46: Επιλογή ID συνόλου δεδομένων



Εικόνα 47: Συμπλήρωση των πληροφοριών του συνόλου δεδομένων



Εικόνα 48: Μήνυμα επιτυχούς μεταφόρτωσης συνόλου δεδομένων



Εικόνα 49: Σύνολο δεδομένων

4.3.3. Δημιουργία και Επικύρωση Μοντέλου

Όπως έχει ήδη αναφερθεί η διαδικασία δημιουργίας ενός μοντέλου στο Jaqpot v5 διαφέρει σημαντικά από την αντίστοιχη στο Jaqpot Quattro και αυτό γιατί ενώ στο Jaqpot Quattro όλα γίνονται από το σύστημα με το χρήστη απλώς να κλικάρει και να επιλέγει αυτά που θέλει, στην νέα εκδοχή ο χρήστης δημιουργεί το μοντέλο του γράφοντας κώδικα σε Python και στην συνέχεια το ανεβάζει στην πλατφόρμα. Και αν το γεγονός της συγγραφής κώδικα μπορεί να κάνει την όλη διαδικασία να φαίνεται περίπλοκη, όπως φαίνεται παρακάτω είναι κάτι παραπάνω από απλή.

Για την καλύτερη κατανόηση του τρόπου δημιουργίας ενός μοντέλου στο Jaqpott v5 θα γίνει παρουσίαση ενός πραγματικού case με την υλοποίηση ενός μοντέλου της βιβλιογραφίας και για λόγους σύγκρισης επιλέγεται να χρησιμοποιηθεί το ίδιο μοντέλο που χρησιμοποιήθηκε και στην παρουσίαση του Jaqpot Quattro και προβλέπει τη διαλυτότητα C60 σε οργανικούς διαλύτες (Gharagheizi, F., & Alamdari, R. F. (2008) "Predicting C60 solubility in organic solvents by means of a molecular-based model).

4.3.3.1. Δημιουργία Μοντέλου

Εισαγωγή της βιβλιοθήκης jaqpotpy και διαφόρων εξαρτημάτων από τη βιβλιοθήκη pandas

```
import pandas as pd

from jaqpotpy import Jaqpot

from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV
```

Διάβασμα του αρχείου csv που περιέχει τα δεδομένα και εκτύπωση των στηλών

```
df=pd.read_csv('70_model.csv') # Reads the data
print(list(df))  # Prints the headers of all columns
```

['Solvents', 'piPC03', 'ATS1m', 'Seigp', 'More23e', 'H1m', 'logS Exp.']

Ορισμός ανεξάρτητων μεταβλητών και εξαρτημένης μεταβλητής

```
Xall=df[['piPC03', 'ATS1m', 'Seigp', 'More23e', 'H1m']] # Defines the columns that will be used as
independent features
Yall=df['logS Exp.']   # Defines the end-point
```

Διαχωρισμός δεδομένων σε training και test dataset

```
X_train, X_test, Y_train, Y_test = train_test_split(Xall, Yall, train_size=0.75, test_size=0.25,
random_state=1)
# Splits the data into training and test sets
```

Ορισμός pipeline που αποτελείται από την προεπεξεργασία ,κλιμάκωση και τον αλγόριθμο πολλαπλής γραμμικής παλινδρόμησης

```
stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]
pipelinelinear = Pipeline(stepslinear) # define the pipeline object.
```

Εκτέλεση διαδικασίας επικύρωσης (5 fold cross validation)

```
cross_val_score(estimator=pipelinelinear, X=X_train, y=Y_train, cv=5, n_jobs=-1)
                                        #Performs a 5-fold cross validation
```

array([0.91906039, 0.88995619, 0.90445436, 0.86506266, 0.62316459])

Εκπαίδευση του μοντέλου

```
pipelinelinear.fit(X_train, Y_train)
print('Training score: ', pipelinelinear.score(X_train, Y_train))
print('Testing score: ', pipelinelinear.score(X_test, Y_test))
print('Total score: ', pipelinelinear.score(Xall, Yall))          #Trains the model and prints R^2 statistics
```

Training score: 0.8994088488271355

Testing score: 0.9043040438111096

Total score: 0.9034772311898356

Ενσωμάτωση μοντέλου στο Jaqpot: Ο χρήστης εισάγει το όνομα χρήστη και τον κωδικό πρόσβασης για να εισέλθει στον λογαριασμό του Jaqpot (ή το api key αν έχει συνδεθεί με λογαριασμό Google ή Github) και με μία μόνο εντολή το ενσωματώνει στο Jaqpot

```
jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")

jaqpot.set_api_key
```

api key is set

```
jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Linear Model for Predicting Solubility of C60 Fullerenes in Various Solvents","Linear Model","linearmodel")
```

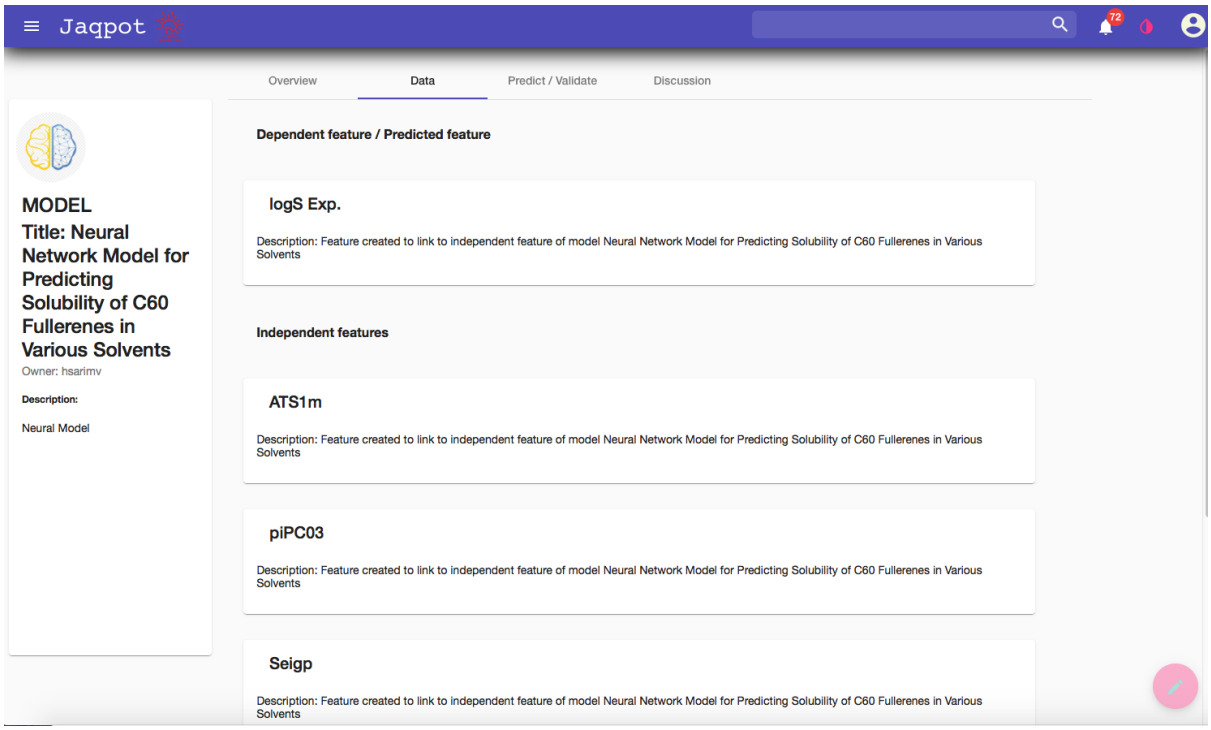
Model with id: Kz6NZU5Aqk5WajFx8OAo created. Please visit https://app.jaqpot.org/

Στην δεύτερη και προαιρετική φάση της μοντελοποίησης, ο χρήστης έχει την δυνατότητα εκμεταλλευόμενος τις δυνατότητες της βιβλιοθήκης Scikit Learn να δημιουργήσει μια αναπαράσταση PMML του μοντέλου η οποία προσφέρει πλήρη αναπαράσταση του μοντέλου, ανεξάρτητη από τη γλώσσα προγραμματισμού παρέχοντας έτσι δυνατότητα πλήρους αναθεώρησης του μοντέλου.

```
from sklearn2pmml.pipeline import PMMLPipeline

pipelinepmmllinear = PMMLPipeline([
  ("scaler", MinMaxScaler()), ("MLR", LinearRegression())
])
pipelinepmmllinear.fit(X_train, Y_train)
```

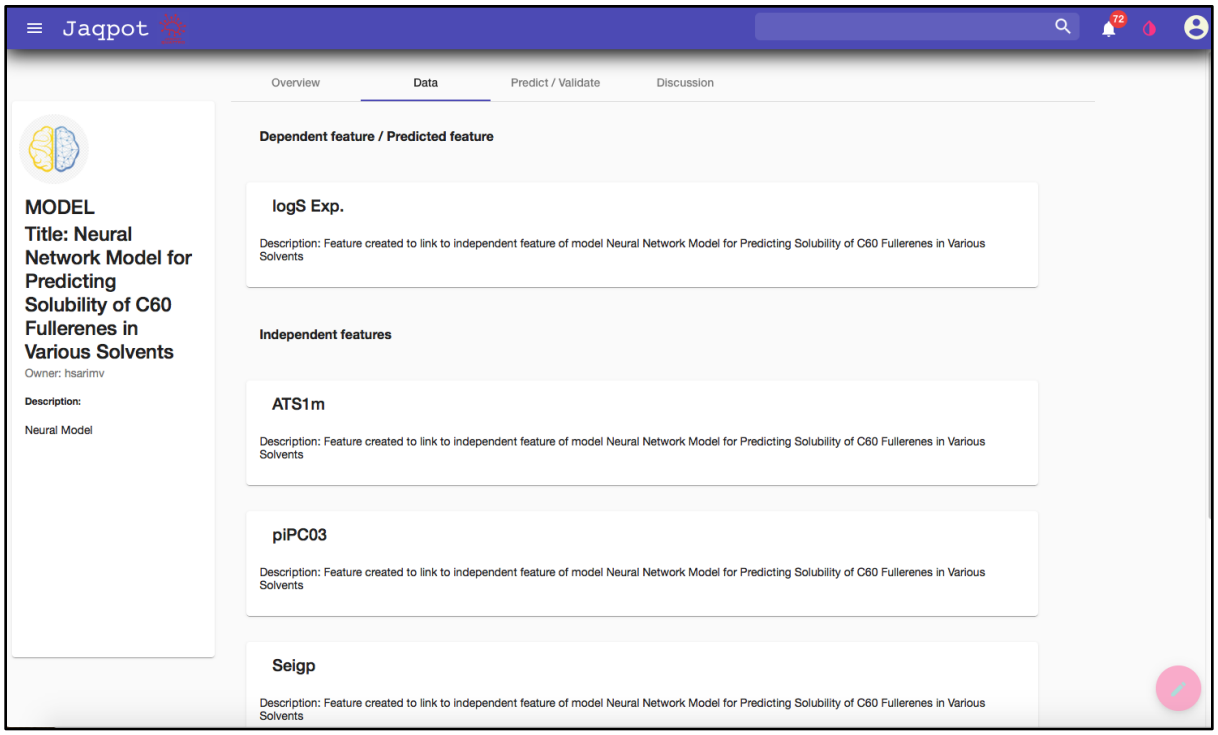
```
from sklearn2pmml import sklearn2pmml

sklearn2pmml(pipelinepmmllinear, "SolubilityC60linear.pmml", with_repr = True)
```

Εάν τα παραπάνω βήματα πραγματοποιηθούν σωστά, τότε δημιουργείται μια διαδικτυακή εκδοχή του μοντέλου στο Jaqpot με ένα μοναδικό αναγνωριστικό URI ενώ υπάρχουν διαθέσιμες και καρτέλες όπου ο χρήστης μπορεί να παρέχει πληροφορίες και λεπτομέρειες σχετικά με το μοντέλο όπως φαίνεται στην Εικόνα 49.



Εικόνα 49: Η σελίδα του μοντέλου με τις διαθέσιμες καρτέλες

Στην καρτέλα Overview παρέχονται οι γενικές πληροφορίες του μοντέλου, οι αναφορές QPRF ενώ μπορεί να προστεθεί και η αναπαράσταση PMML. Στην καρτέλα Data ο χρήστης μπορεί να παράσχει συγκεκριμένες πληροφορίες σχετικά με τις ανεξάρτητες αλλά και την εξαρτημένη μεταβλητή του μοντέλου όπως περιγραφές, μονάδες και οντολογικές τάξεις όπως φαίνεται στην Εικόνα 50.

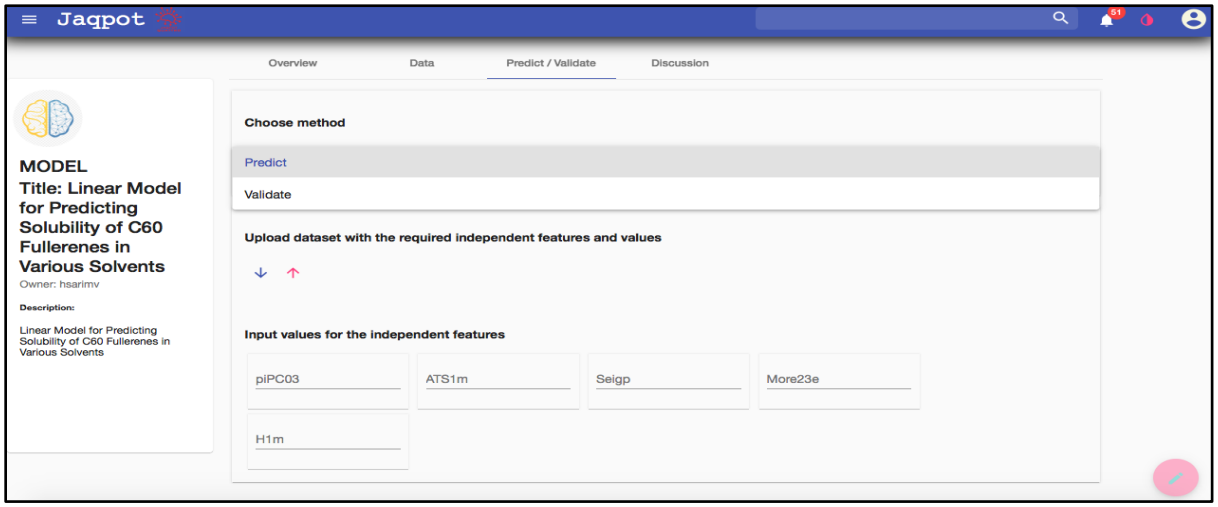


Εικόνα 50: Καρτέλα Data

4.3.3.2. Χρήση Μοντέλου για πρόβλεψη

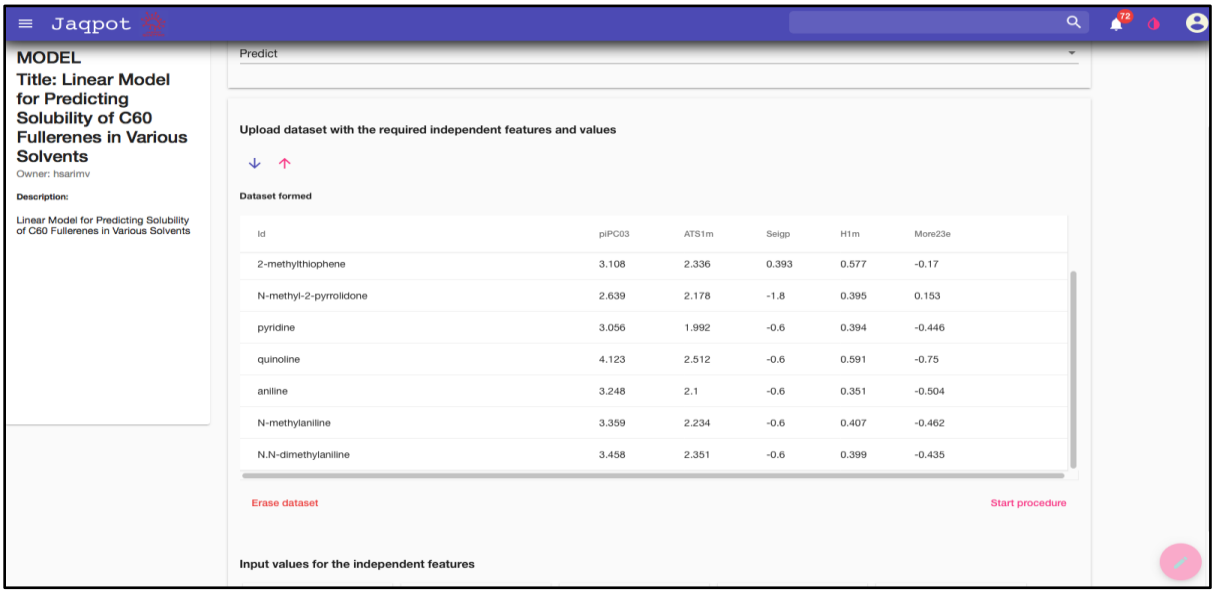
Η καρτέλα Predict/Validate αποτελεί την σημαντικότερη καρτέλα αφού είναι αυτή με τις κυριότερες λειτουργίες του μοντέλου (Εικόνα 51). Μέσω αυτής ο χρήστης έχει την δυνατότητα να κάνει προβλέψεις γνωρίζοντας τις τιμές των ανεξάρτητων μεταβλητών αλλά αγνοώντας την τιμή της εξαρτημένης αλλά και να τεστάρει το μοντέλο χρησιμοποιώντας ένα ολοκληρωμένο σύνολο δεδομένων (συμπεριλαμβανομένων και των τιμών της εξαρτημένης μεταβλητής). Για την εισαγωγή των δεδομένων παρέχονται από το σύστημα δύο επιλογές. Αυτή της χειροκίνητης εισαγωγής, όταν πρόκειται για σχετικά μικρά σύνολα δεδομένων, και αυτή της μεταφόρτωσης χρησιμοποιώντας πρότυπα csv που δημιουργούνται αυτόματα για κάθε μοντέλο από το σύστημα και μπορούν να μεταφορτωθούν για να συμπληρωθούν με δεδομένα (κάνοντας κλικ στο μπλε κάτω βέλος). Τα πρότυπα περιέχουν όλα τα ονόματα μεταβλητών εισόδου, οπότε ο χρήστης μπορεί να

προσθέσει τις αντίστοιχες τιμές σε κάθε στήλη και να μεταφορτώσει τα δεδομένα κάνοντας κλικ στο κόκκινο βέλος που δείχνει προς τα πάνω (Εικόνα 51).

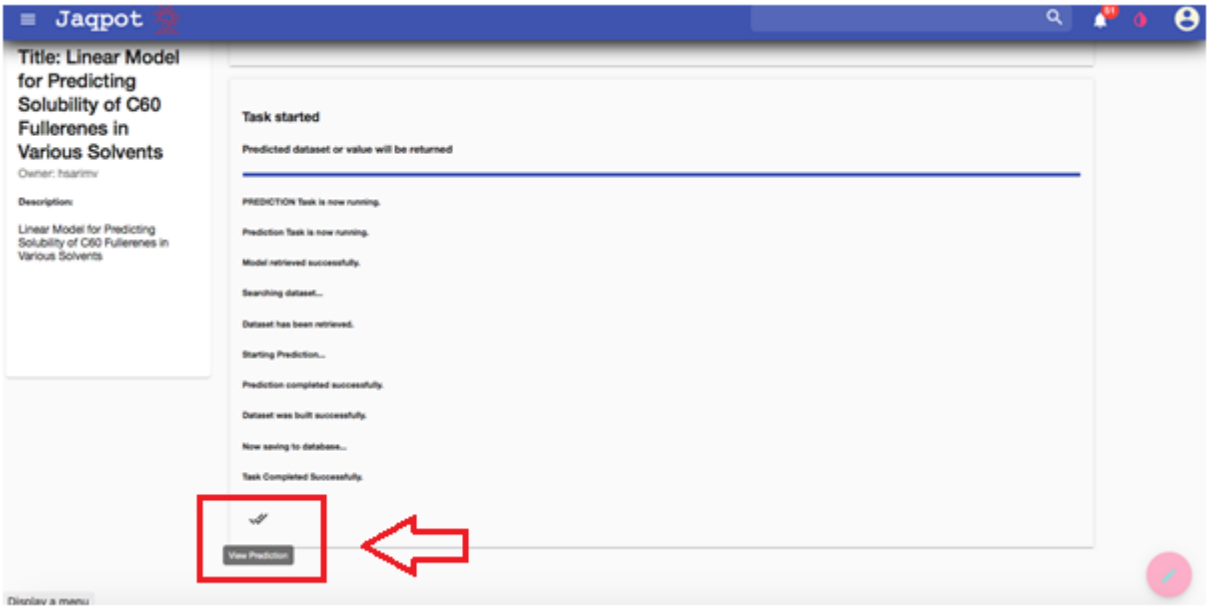


Εικόνα 51: Καρτέλα Predict/Validate

Αφού ολοκληρωθεί η διαδικασία μεταφόρτωσης του συνόλου δεδομένων και ο ορισμός της στήλης που λειτουργεί ως αναγνωριστικό (ID) εμφανίζεται μια προεπισκόπηση του συνόλου δεδομένων (Εικόνα 52) και κλικάροντας το «Start procedure» εκκινεί η διαδικασία της πρόβλεψης, ενώ κλικάροντας το «View Prediction» που φαίνεται στην Εικόνα 53 μπορούμε να δούμε τις προβλέψεις που πραγματοποιήθηκαν.



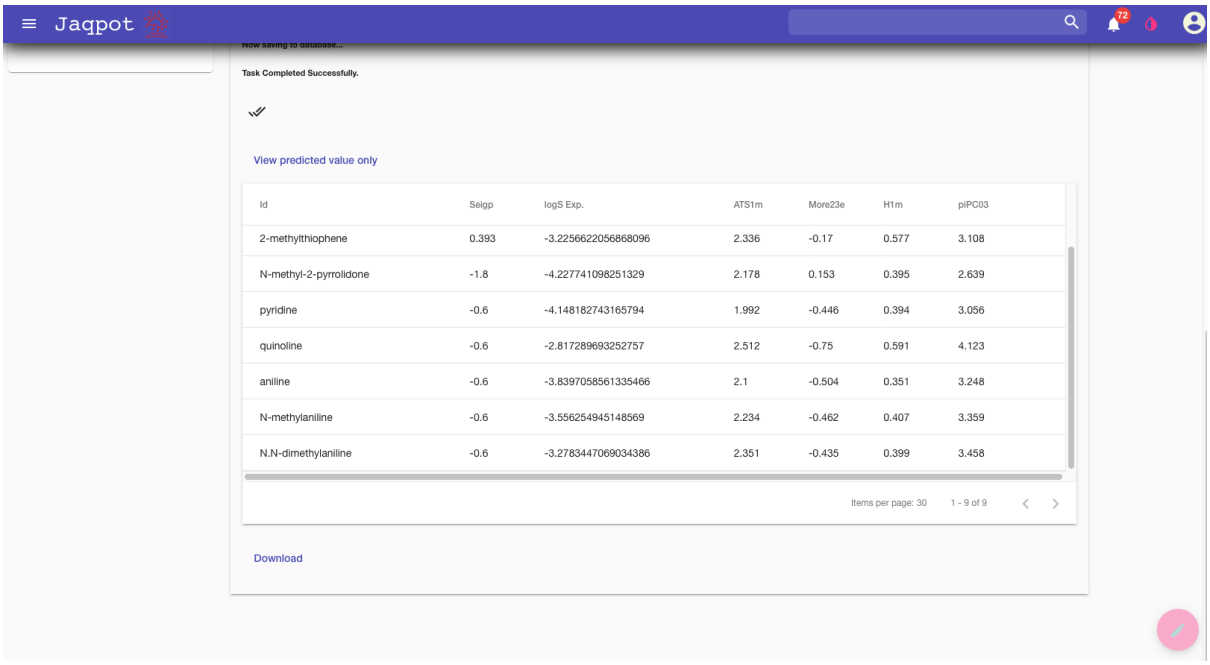
Εικόνα 52: Καρτέλα Predict/Validate



Εικόνα 53: Καρτέλα Predict/Validate

4.3.3.3. Επικύρωση Μοντέλου

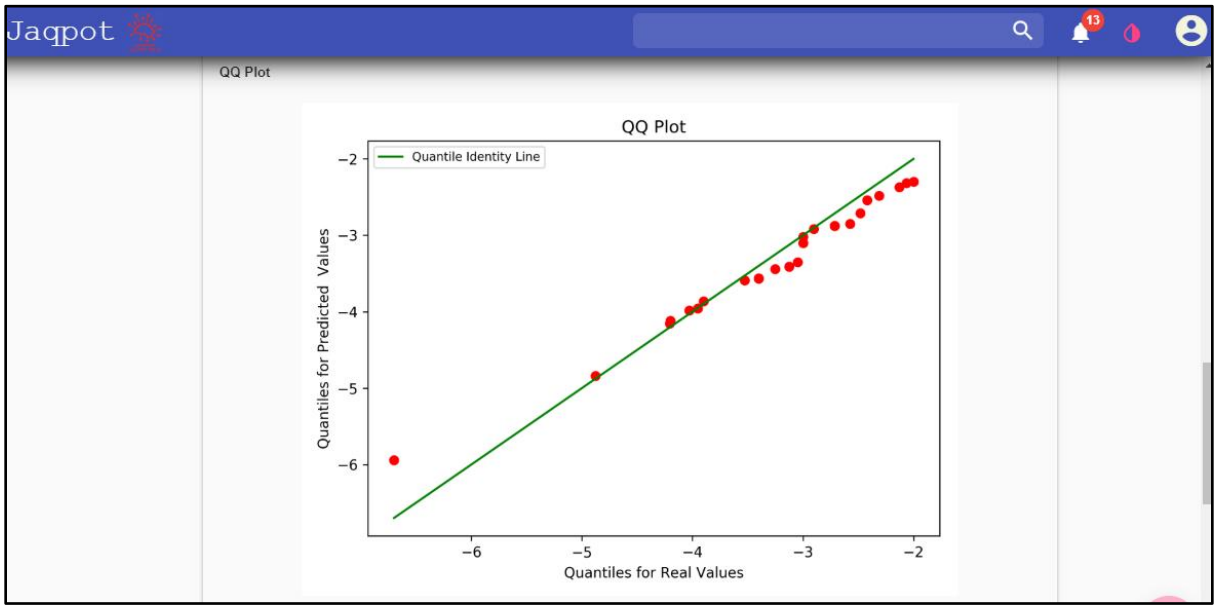
Για την επικύρωση του μοντέλου επιλέγουμε την επιλογή «Validate» στην οθόνη της Εικόνας 51 και η διαδικασία της επικύρωσης οδηγεί στην παραγωγή μιας αναφοράς με στατιστικά στοιχεία επικύρωσης, συνοδευόμενη από μια γραφική παράσταση QQ και γραφική παράσταση Real vs. Predicted, δίνοντας μια εικόνα για την αποτελεσματικότητα του μοντέλου όπως φαίνεται στις Εικόνες 54-56.



Id	Seigp	logS Exp.	ATS1m	More23e	H1m	piPC03
2-methylthiophene	0.393	-3.2256622056868096	2.336	-0.17	0.577	3.108
N-methyl-2-pyrrolidone	-1.8	-4.227741098251329	2.178	0.153	0.395	2.639
pyridine	-0.6	-4.148182743165794	1.992	-0.446	0.394	3.056
quinoline	-0.6	-2.817289693252757	2.512	-0.75	0.591	4.123
aniline	-0.6	-3.8397058561335466	2.1	-0.504	0.351	3.248
N-methylaniline	-0.6	-3.556254945148569	2.234	-0.462	0.407	3.359
N,N-dimethylaniline	-0.6	-3.2783447069034386	2.351	-0.435	0.399	3.458

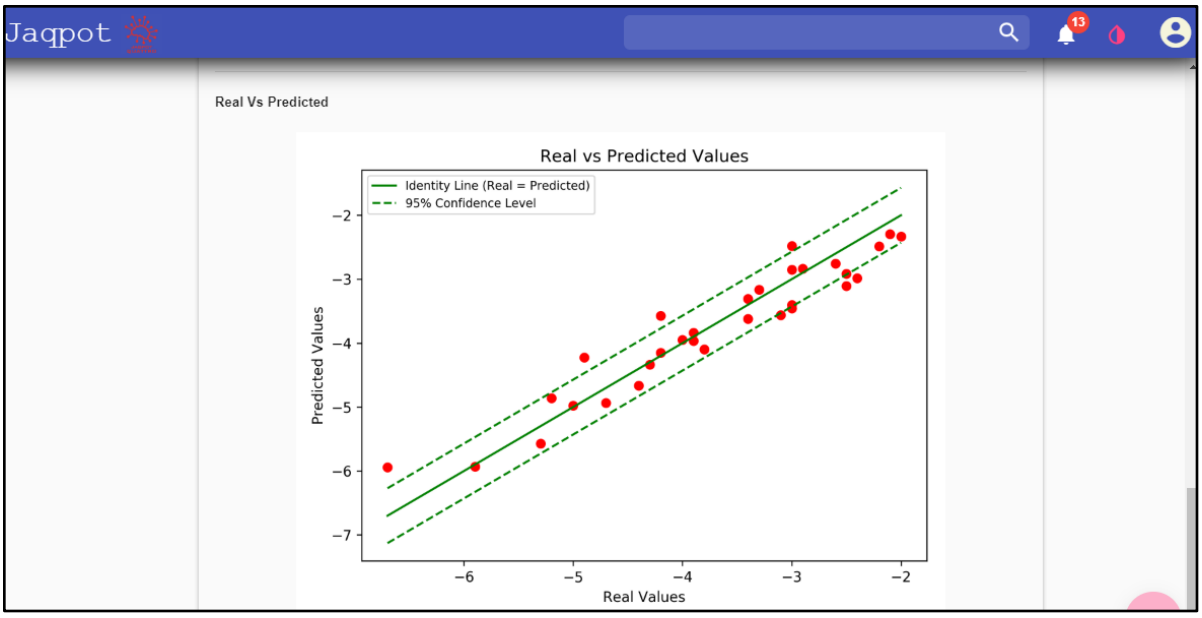
Items per page: 30 1 - 9 of 9

Εικόνα 54: Προβλεπόμενες τιμές μετά τη μοντελοποίηση



Εικόνα 55: Αποτελέσματα επικύρωσης - Σχέδιο QQ (Jaqpot 5).

Source Information	Model URI



Εικόνα 56:Αποτελέσματα επικύρωσης - Real vs. Predicted plot (Jaqpot 5).

4.4. Μοντέλα που υλοποιήθηκαν

Μετά την αναλυτική παρουσίαση των δύο εργαλείων γίνεται εύκολα αντιληπτό πως οι δύο εκδόσεις του Jaqpot προσφέρουν πολύ σημαντικές δυνατότητες όσον αφορά την δημιουργία, την αποθήκευση, την επικύρωση και την χρήση των μοντέλων για πρόβλεψη. Σε αυτό το πλαίσιο προέκυψε η ιδέα για την δημιουργία ενός διαδικτυακού αποθετηρίου μοντέλων QSAR το οποίο θα είναι διαθέσιμο σαν web service και θα είναι εύκολο να μοιραστεί με την ερευνητική κοινότητα.

Τα μοντέλα που υλοποιήθηκαν, είναι μοντέλα της βιβλιογραφίας και επιλέχθηκαν μετα από βιβλιογραφική έρευνα με βασικότερο κριτήριο την διαθεσιμότητα των δεδομένων που θα επέτρεπαν την δημιουργία του μοντέλου. Τα μοντέλα που υλοποιήθηκαν παρουσιάζονται στον παρακάτω πίνακα όπου συμπεριλαμβάνεται και το URI του μοντέλου σε καθένα από τα δύο εργαλεία.

Model name	Year	Endpoint	Jaqpot Quattro	Jaqpot v5
Methodology for developing structure-activity evaluation to identify combinations of physical features of nanomaterial that influence potential cell damage by MLR/LDA (TiO2 case)	2010	In vitro - Cytotoxicity - membrane damage measured as lactate dehydrogenase (LDH) release [units/L]	http://jaqpot.org/m_detail?name=2KoKHcIgMJloSeWuZ03a	https://app.jaqpot.org/model/izMncmc5LMbgC6o7Fkj8
Methodology for developing structure-activity evaluation to identify combinations of physical features of nanomaterial that influence potential cell damage by MLR/LDA (ZnO case)	2010	In vitro - Cytotoxicity - membrane damage measured as lactate dehydrogenase (LDH) release [units/L]	http://jaqpot.org/m_detail?name=cm49KGhUjkMw6wyntKQF	https://app.jaqpot.org/model/fvAe4KnIOOiNgGf7Ve1p
Regression model to understand the aggregated ZVCN against E.Coli by MLR (Placket-Burman design)	2010	In vitro - Cytotoxicity - measured as percentage of dead E. Coli population	http://jaqpot.org/m_detail?name=48JYATz0KFTkZjGd8AfS	https://app.jaqpot.org/model/8su6n4cfcJpzZD2NDZGN
Prediction of the Biological surface adsorption index (BSAI) on different NPs by MLR	2011	log(k) k: adsorption coefficient	http://jaqpot.org/m_detail?name=DjRQk8AqG42nckg5KoxZ	https://app.jaqpot.org/model/gSvjUZ17EEAV5OWL7Uls
Predictive model of TiO2 NPs damage on membrane cell by SMILES-based optimal descriptor and Monte Carlo technique (CORAL software)	2013	In vitro - Cytotoxicity - membrane damage measured as lactate dehydrogenase (LDH) release [units/L]	http://jaqpot.org/m_detail?name=4oxlwXBZMJ4suYFTSl4d	https://app.jaqpot.org/model/nTJgb4Ss3zHIYZEcbg78
Cytotoxicity of metal oxide to bacteria E.Coli models by Periodic table-based descriptors and stepwise-MLR	2014	In vitro - Cytotoxicity - measured as pEC50	http://jaqpot.org/m_detail?name=EFffilLYKMgLUq3qNYBw	https://app.jaqpot.org/model/QgRRwyU8r7e0NubEuDdX
Photo-induced toxicity of metal oxide NPs to E. Coli by MLR (dark condition case)	2014	In vitro - Cytotoxicity - measured as -log(LC50)	http://jaqpot.org/m_detail?name=KIWUeelVM8x7x1iC7cXi	https://app.jaqpot.org/model/hygpzrH71XS1Wr8lGS69
Photo-induced toxicity of metal oxide NPs to E. Coli by MLR (Photo-induced (light) case)	2014	In vitro - Cytotoxicity - measured as -log(LC50)	http://jaqpot.org/m_detail?name=o6Jr81BfQtUddgmwqaee	https://app.jaqpot.org/model/5gCY316DzDh1Fdw4aigo
Predicting metal oxide Nps toxicity to E. Coli cell line by MLR	2016	In vitro - Cytotoxicity - measured as log(1/EC50)	http://jaqpot.org/m_detail?name=qul6HlLHSypXWX8zvMQ3	https://app.jaqpot.org/model/OAiBYuee5PLJ7F580f2J
Predicting C60 solubility in organic solvents by SMILES-based optimal descriptor and Monte Carlo technique	2008	Solubility in organic solvents	http://jaqpot.org/m_detail?name=sCoqY3D3xCpSuyS6RdoQ	https://app.jaqpot.org/model/VRp8f6A4DuJc8fsavvpB

Methodology for developing structure-activity evaluation to identify combinations of physical features of nanomaterial that influence potential cell damage by MLR/LDA (TiO2 case)	
Πηγή:	Sayes, C., & Ivanov, I. (2010). Comparative Study of Predictive Computational Models for Nanoparticle-Induced Cytotoxicity. Risk Analysis, 30(11), 1723–1734. (TiO2 case) http://doi.org/10.1111/j.1539-6924.2010.01438.x
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - membrane damage measured as lactate dehydrogenase (LDH) release [units/L]
Μεταβλητές Εισόδου (Input Variables)	Size in water Concentration [mg/L] Zeta Potential [mV]
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=OVRXa54JPk1iSk Metal Oxide: TiO2 24 measures, combination of: Engineered Size (30, 45, 125) 2 x Concentration(25, 50, 100, 200)
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=2KoKHcIgMJloSeWuZ03a
Jaqpot v5:	https://app.jaqpot.org/model/izMncmc5LMbgC6o7Fkj8

Methodology for developing structure-activity evaluation to identify combinations of physical features of nanomaterial that influence potential cell damage by MLR/LDA (ZnO case)	
Πηγή:	Sayes, C., & Ivanov, I. (2010). Comparative Study of Predictive Computational Models for Nanoparticle-Induced Cytotoxicity. Risk Analysis, 30(11), 1723–1734. (ZnO case) http://doi.org/10.1111/j.1539-6924.2010.01438.x
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - membrane damage measured as lactate dehydrogenase (LDH) release [units/L]
Μεταβλητές Εισόδου (Input Variables)	Size in water Concentration [mg/L] Size in CCM
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=OVRXa54JPk1iSk Metal Oxide: ZnO 18 measures, combination of: Engineered Size (50, 60, 70, 1000, 1200, 1500) Concentration(25, 50, 100)
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=cm49KGhUjkMw6wyntKQF
Jaqpot v5:	https://app.jaqpot.org/model/fvAe4KnIOOiNgGf7Ve1p

Regression model to understand the aggregated ZVCN against E.Coli by MLR (Placket-Burman design)	
Πηγή:	Rispoli, F., Angelov, A., Badia, D., Kumar, A., Seal, S., & Shah, V. (2010). Understanding the toxicity of aggregated zero valent copper nanoparticles against Escherichia coli. Journal of Hazardous Materials, 180(1-3), 212–216. http://doi.org/10.1016/j.jhazmat.2010.04.016
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - measured as percentage of dead E. Coli population
Μεταβλητές Εισόδου (Input Variables)	pH Temperature Aertion rate Concentration of nanoparticles Concentration of bacteria
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=7Sra2sRbK1lJRt 16 Metal ZVCN: zero valent cooper Cu nanoparticle
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=48JYATz0KFTkZjGd8AfS
Jaqpot v5:	https://app.jaqpot.org/model/8su6n4cfcJpzZD2NDZGN

Prediction of the Biological surface adsorption index (BSAI) on different NPs by MLR	
Πηγή:	"Xia, X. R., Monteiro-Riviere, N. A., Mathur, S., Song, X., Xiao, L., Oldenberg, S. J., … Riviere, J. E. (2011). Mapping the surface adsorption forces of nanomaterials in biological systems. ACS Nano, 5(11) , 9074-9081 Chen, R., Zhang, Y., Monteiro-Riviere, N. A., & Riviere, J. E. (2016). Quantification of nanoparticle pesticide adsorption: computational approaches based on experimental data. Nanotoxicology, 10(8), 1118–1128. http://doi.org/10.1021/nn203303c
Εξαρτημένη Μεταβλητή (Endpoint)	log(k) k: adsorption coefficient
Μεταβλητές Εισόδου (Input Variables)	V: Lipophilicity interaction β: Hydrogenbond basicity α: Hydrogenbond acidity π: Pipolarity/polarizability R: lone-pair electrons
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=lPIrs0YslSompB 28 Carbon-based MWCNT40nm-COOH
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=DjRQk8AqG42nckg5KoxZ
Jaqpot v5:	https://app.jaqpot.org/model/gSvjUZ17EEAV5OWL7Uls

Predictive model of TiO2 NPs damage on membrane cell by SMILES-based optimal descriptor and Monte Carlo technique (CORAL software)	
Πηγή:	Toropova, A. P., & Toropov, A. A. (2013). Optimal descriptor as a translator of eclectic information into the prediction of membrane damage by means of various TiO2 nanoparticles. Chemosphere, 93(10), 2650–2655. Toropova, A. P., Toropov, A. A., Benfenati, E., Puzyn, T., Leszczynska, D., & Leszczynski, J. (2014). Optimal descriptor as a translator of eclectic information into the prediction of membrane damage: The case of a group of ZnO and TiO2 nanoparticles. Ecotoxicology and Environmental Safety, 108, 203–209. http://doi.org/10.1016/j.chemosphere.2013.09.089
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - membrane damage measured as lactate dehydrogenase (LDH) release [units/L]
Μεταβλητές Εισόδου (Input Variables)	Engineered Size Size in water Size in PBS Concentration Zeta potential
Σύνολο Δεδομένων (Dataset)	http://www.jaqpot.org/data_detail?name=yW3pfohTS3l33m&page=2 TiO2 Metal Oxide Engineered Size: 30, 45, 125 Size in water: 101-967 Size in PBS: 961-3871/
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=4oxlwXBZMJ4suYFTSl4d
Jaqpot v5:	https://app.jaqpot.org/model/nTJgb4Ss3zHIYZEcbg78

Cytotoxicity of metal oxide to bacteria E.Coli models by Periodic table-based descriptors and stepwise-MLR	
Πηγή:	Kar, S., Gajewicz, A., Puzyn, T., Roy, K., & Leszczynski, J. (2014). Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: A mechanistic QSTR approach. Ecotoxicology and Environmental Safety, 107, 162–169. http://doi.org/10.1016/j.ecoenv.2014.05.026
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - measured as pEC50
Μεταβλητές Εισόδου (Input Variables)	Xox: charge of metal cation corresponding to a given oxide
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=7E6XB8VzAGEZNl 17 Metal Oxides ZnO CuO Al2O3 Fe2O3 SnO2 TiO2 V2O3 Y2O3 Bi2O3 In2O3 Sb2O3 SiO2 ZrO2 CoO NiO Cr2O3 La2O3
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=EFffilLYKMgLUq3qNYBw
Jaqpot v5:	https://app.jaqpot.org/model/QgRRwyU8r7e0NubEuDdX

Photo-induced toxicity of metal oxide NPs to E. Coli by MLR (dark condition case)	
Πηγή:	Pathakoti, K., Huang, M.-J., Watts, J. D., He, X., & Hwang, H.-M. (2014). Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. Journal of Photochemistry and Photobiology B: Biology, 130, 234–240. (dark condition case) http://doi.org/10.1016/j.jphotobiol.2013.11.023
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - measured as -log(LC50)
Μεταβλητές Εισόδου (Input Variables)	MELECT: the absolute electronegativity of the metal atom LZELEHHO: the absolute electronegativity of the metal oxide
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=yjBO04fO3d19XL 17 Metal Oxides ZnO CuO Al2O3 Fe2O3 SnO2 TiO2 V2O3 Y2O3 Bi2O3 In2O3 Sb2O3 SiO2 ZrO2 CoO NiO Cr2O3 La2O3
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=KIWUeelVM8x7x1iC7cXi
Jaqpot v5:	https://app.jaqpot.org/model/hygpzrH71XS1Wr8lGS69

Photo-induced toxicity of metal oxide NPs to E. Coli by MLR(Photo-induced (light) case)	
Πηγή:	Pathakoti, K., Huang, M.-J., Watts, J. D., He, X., & Hwang, H.-M. (2014). Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. Journal of Photochemistry and Photobiology B: Biology, 130, 234–240. (Photo-induced (light) case) http://doi.org/10.1016/j.jphotobiol.2013.11.023
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - measured as -log(LC50)
Μεταβλητές Εισόδου (Input Variables)	Cp is the literature molar heat capacity of the metal oxide at 298.15 K. ALZLUMO is the average of the alpha and beta LUMO energies of the metal oxide.
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=4xoqetJXfkMB0S 17 Metal Oxides ZnO CuO Al2O3 Fe2O3 SnO2 TiO2 V2O3 Y2O3 Bi2O3 In2O3 Sb2O3 SiO2 ZrO2 CoO NiO Cr2O3 La2O3
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=o6Jr81BfQtUddgmwqaee
Jaqpot v5:	https://app.jaqpot.org/model/5gCY316DzDh1Fdw4aigo

Predicting metal oxide Nps toxicity to E. Coli cell line by MLR	
Πηγή:	Mu, Y., Wu, F., Zhao, Q., Ji, R., Qie, Y., Zhou, Y., … Xing, B. (2016). Predicting toxic potencies of metal oxide nanoparticles by means of nano-QSARs. Nanotoxicology. State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing, China. http://doi.org/10.1080/17435390.2016.1202352
Εξαρτημένη Μεταβλητή (Endpoint)	In vitro - Cytotoxicity - measured as log(1/EC50)
Μεταβλητές Εισόδου (Input Variables)	Z/r : Polarization force parameter ΔH_Me+ : represents the enthalpy of formation of a gaseous cation having the same oxidation state as that in the metal oxide structure.
Σύνολο Δεδομένων (Dataset)	http://jaqpot.org/data_detail?name=4fk3ZAJrRhskX4 17 Metal Oxides ZnO CuO Al2O3 Fe2O3 SnO2 TiO2 V2O3 Y2O3 Bi2O3 In2O3 Sb2O3 SiO2 ZrO2 CoO NiO Cr2O3 La2O3
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=qul6HlLHSypXWX8zvMQ3
Jaqpot v5:	https://app.jaqpot.org/model/OAiBYuee5PLJ7F580f2J

Predicting C60 solubility in organic solvents by SMILES-based optimal descriptor and Monte Carlo technique	
Πηγή:	Garagheizi, F., & Alamdari, R. F. (2008). A molecular-based model for prediction of solubility of C60 fullerene in various solvents. Fullerenes Nanotubes and Carbon Nanostructures, 16(1), 40–57. http://doi.org/10.1080/15363830701779315
Εξαρτημένη Μεταβλητή (Endpoint)	Solubility in organic solvents- measured as logS Exp
Μεταβλητές Εισόδου (Input Variables)	Molecular descriptors are defined for solvents according to chemical structure using the Dragon Software. piPC03: Molecular multiple path count of order 03 (walk and path counts) ATS1m 2D: Broto-Mreau autocorrelation of a topological structure-lag 1/weighted by atomic masses (2D autocorrelations) Seigp: Eigenvalue sum from polarizability weighted distance matrix (Eigenvalue 0 based indices) More23e: 3D-MORSE-signal 23/weighted by atomic sanderson electronegativitie (More23e 3D-MORSE descriptors) H1m: H autocorrelation of lag 1/weighted by atomic masses (GETAWAY descriptors)"
Σύνολο Δεδομένων (Dataset)	Training: http://www.jaqpot.org/data_detail?name=XmCQVC7o5jKKRv Test http://www.jaqpot.org/data_detail?name=3UbgEJPIdT2Ovs Carbon-based NM, Fullerene C60
Αλγόριθμος	**Multiple Linear Regression**
Jaqpot Quattro:	http://jaqpot.org/m_detail?name=sCoqY3D3xCpSuyS6RdoQ
Jaqpot v5:	https://app.jaqpot.org/model/VRp8f6A4DuJc8fsavvpB

5. Συμπεράσματα & Προτάσεις για Μελλοντική Έρευνα

Φτάνοντας στο τέλος αυτής της εργασίας θα μπορούσαμε να πούμε πως έχει γίνει αντιληπτή η ανάγκη για τη μοντελοποίηση της τοξικότητας των κατασκευασμένων νανοϋλικών, μιας και αυτά χρησιμοποιούνται όλο και περισσότερο στην καθημερινή μας ζωή. Η υπολογιστική προσέγγιση είναι μη χρονοβόρα διαδικασία με χαμηλότερο κόστος που συμπληρώνει τις παραδοσιακές πειραματικές μεθόδους, συμβάλει στην ασφαλέστερη χρήση της νανοτεχνολογίας, ενώ μπορεί να χρησιμοποιηθεί τόσο από τις αρχές για τον έλεγχο και την πρόληψη της έκθεσης σε τοξικά νανοϋλικά όσο και από την βιομηχανική κοινότητα για το σχεδιασμό νανοϋλικών, που έχουν τα επιθυμητά χαρακτηριστικά, αλλά ταυτόχρονα ικανοποιούν τις προδιαγραφές ασφάλειας.

Προς αυτή την κατεύθυνση κινήθηκε η ανάπτυξη των δύο πλατφορμών ανοιχτού κώδικα Jaqpot Quattro και Jaqpot v5, φιλοδοξώντας να παρέχουν στον χρήστη προηγμένες δυνατότητες μοντελοποίησης. Πράγματι μέσω των δύο αυτών διαδικτυακών εφαρμογών, οι χρήστες έχουν τη δυνατότητα να εργαστούν με ποικιλία πληροφοριών, αφού μπορούν να αναζητήσουν ανάμεσα στα ήδη υπάρχοντα αποθηκευμένα σετ δεδομένων ή να ανεβάσουν τα δικά τους δεδομένα. Αντίστοιχα, οι χρήστες μπορούν να χρησιμοποιήσουν τα ήδη διαθέσιμα μοντέλα ή να δημιουργήσουν και να μοιραστούν με την κοινότητα νέα μοντέλα, επιλέγοντας από μια μεγάλη βιβλιοθήκη αλγορίθμων (για το Jaqpot Quattro) ή γράφοντας λίγες γραμμές κώδικα στη γλώσσα Python (για το Jaqpot v5). Και στις δύο περιπτώσεις ο χρήστης έχει την δυνατότητα να αξιολογήσει το μοντέλο που παράγεται, ανάμεσα σε διαφορετικές τεχνικές επικύρωσης (cross validation, split validation, external validation).

Στην παρούσα εργασία αναπτύχθηκε μια βιβλιοθήκη μοντέλων που είναι διαθέσιμη στην ερευνητική κοινότητα για πραγματοποίηση προβλέψεων, αξιολόγηση και περαιτέρω βελτίωση και ανάπτυξη. Τα μοντέλα που υλοποιήθηκαν με την χρήση αλγορίθμων της μηχανικής μάθησης χρησιμοποιούνται για την πρόβλεψη διαφόρων ιδιοτήτων νανοϋλικών. Στην πλειοψηφία τους προβλέπουν τοξικότητα εκφρασμένη είτε ως ποσοστό νεκρού πληθυσμού του βακτηρίου Escherichia coli, είτε ως απελευθέρωση του ενζύμου γαλακτική αφυδρογονάση (LDH) που επιφέρει βλάβες στην κυτταρική μεμβράνη ενώ υπάρχουν και μοντέλα που προβλέπουν άλλες φυσικοχημικές ιδιότητες όπως ο λογάριθμος της σταθεράς προσρόφησης ή η διαλυτότητα του φουλερενίου σε οργανικούς

διαλύτες. Τα δύο εργαλεία ανταποκρίθηκαν πολύ καλά στις απαιτήσεις τις υπολογιστικής νανοτοξικολογίας και τα παραγόμενα μοντέλα όχι μόνο είναι εύκολα ως προς τη χρήση του αλλά παρέχουν πλούσιες πληροφορίες σχετικά με τους αλγόριθμους μοντελοποίησης, τις μεταβλητές που χρησιμοιούνται ως μεταβλητές εισόδου και εξόδου, τις διαδικασίες εξιολόγησης και το χώρο στον οποίο οι προβλέψεις θεωρούνται αξιόπιστες.

Η ανοιχτού κώδικα φύση των εργαλείων καθιστά ευκολότερη την προσθήκη νέων λειτουργιών μελλοντικά. Δεδομένου του ότι όπως ήδη προαναφέρθηκε στον τομέα της τοξικολογίας εμπλέκονται επιστήμονες πολλών κλάδων, κύριος άξονας της περαιτέρω ανάπτυξης των εργαλείων θα πρέπει να είναι η προώθηση της διεπιστημονικής συνεργασίας (π.χ. μεταξύ των ερευνητών του πειραματικού τομέα και των ερευνητών που ειδικεύονται σε υπολογιστικές προσεγγίσεις). Κλείνοντας θα μπορούσαμε να πούμε με πως οι δύο εφαρμογές ανοιχτού κώδικα (Jaqpot Quattro και Jaqpot v5) μπορούν να συμβάλλουν στην εξέλιξη της υπολογιστικής νανοτοξικολογίας και να αποτελέσουν σημαντικό εργαλείο για τους επιστήμονες που δραστηριοποιούνται στον αναδυόμενο και πολύ ελπιδοφόρο επιστημονικό τομέα της νανοτεχνολογίας. Η παρούσα εργασία αποτελεί την πρώτη προσπάθεια δημιουργίας μιας διαδικτυακής βιβλιοθήλης προβλεπτικών μοντέλων, με την προοπτική αυτή να επεκταθεί και να αποτελέσει κεντρικό σημείο συνεργασίας και αλληλεπίδρασης μεταξύ των ερευνητικών ομάδων.

6. Παράρτημα

Methodology for developing structure-activity evaluation to identify combinations of physical features of nanomaterial that influence potential cell damage by MLR/LDA (TiO2 case)

https://app.jaqpot.org/model/izMncmc5LMbgC6o7Fkj8

```
import pandas as pd
from jaqpotpy import Jaqpot

from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV

df = pd.read_csv('2a_dataset_regression.csv')

print(list(df))   # Prints the headers of all column

Xall=df[['Size in Water (nm)', 'Conc. (mg/L)', 'Zeta Potential (mV)']] # Define the co
lumns that will be used as independent features

Yall=df['Membrane Damage (units/L)']    # Define the end-point

stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]

pipelinelinear = Pipeline(stepslinear) # define the pipeline object.

pipelinelinear.fit(Xall, Yall)
print('Total score: ', pipelinelinear.score(Xall, Yall)  #Trains the model and prints
R^2 statistics
jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")
jaqpot.set_api_key("……")
jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting LDH After TiO2 Cellu
lar Exposures","Model developed by Sayes et al in 2010","linearmodel")
```

Methodology for developing structure-activity evaluation to identify combinations of physical features of nanomaterial that influence potential cell damage by MLR/LDA (ZnO case)

https://app.jaqpot.org/model/fvAe4KnIOOiNgGf7Ve1p

```
import pandas as pd
from jaqpotpy import Jaqpot
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV


df = pd.read_csv('2b_dataset_regression.csv')


print(list(df))   # Prints the headers of all column


Xall=df[['Size in Water', 'Size in CCM', 'Conc.']] # Define the columns that will be u
sed as independent features
Yall=df['Membrane Damage']    # Define the end-point


stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]


pipelinelinear = Pipeline(stepslinear) # define the pipeline object.


pipelinelinear.fit(Xall, Yall)
print('Total score: ', pipelinelinear.score(Xall, Yall))   #Trains the model and print
s R^2 statistics


jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")


jaqpot.set_api_key("……")
jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting LDH After ZnO Cellul
ar Exposures","Model developed by Sayes et al in 2010","linearmodel")
```

```
Model with id: fvAe4KnIOOiNgGf7Ve1p created. Please visit https://app.jaqpot.org/
```

Regression model to understand the aggregated ZVCN against E.Coli by MLR (Placket-Burman design)

https://app.jaqpot.org/model/8su6n4cfcJpzZD2NDZGN

```
import pandas as pd
from jaqpotpy import Jaqpot
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV


df = pd.read_csv('4b_dataset.csv')


print(list(df))   # Prints the headers of all column


Xall=df[['pH', 'Temp', 'Aeration', '[Nano]', '[E. coli]']] # Define the columns that w
ill be used as independent features


Yall=df['Toxicity (%)']    # Define the end-point


stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]


pipelinelinear = Pipeline(stepslinear) # define the pipeline object.


pipelinelinear.fit(Xall, Yall)


print('Total score: ', pipelinelinear.score(Xall, Yall))        #Trains the model and
prints R^2 statistics


jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")


#edo
jaqpot.set_api_key("……")


jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting cytotoxicity for cop
per NPs, Placket-Burman design","Model developed by Rispoli et al in 2010","linearmode
l")
```

```
Model with id: 8su6n4cfcJpzZD2NDZGN created. Please visit https://app.jaqpot.org/
```

Prediction of the Biological surface adsorption index (BSAI) on different NPs by MLR
https://app.jaqpot.org/model/gSvjUZ17EEAV5OWL7Uls

```
import pandas as pd
from jaqpotpy import Jaqpot

from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV


df = pd.read_csv('7_dataset.csv')

print(list(df))   # Prints the headers of all column

Xall=df[['R', 'p', 'a', 'b', 'V']] # Define the columns that will be used as independe
nt features


Yall=df['Log k']    # Define the end-point


X_train, X_test, Y_train, Y_test = train_test_split(Xall, Yall,
                                    train_size=0.75, test_size=0.25, random_state=1)
# Splits the data into training and test sets

stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]


pipelinelinear = Pipeline(stepslinear) # define the pipeline object.


cross_val_score(estimator=pipelinelinear, X=X_train, y=Y_train, cv=5, n_jobs=-1)  #Per
froms a 5-fold cross validation


pipelinelinear.fit(X_train, Y_train)
print('Training score: ', pipelinelinear.score(X_train, Y_train))
print('Testing score: ', pipelinelinear.score(X_test, Y_test))
print('Total score: ', pipelinelinear.score(Xall, Yall))                    #Trains t
he model and prints R^2 statistics


jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")


jaqpot.set_api_key("…….")
```

```
jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting log(k) in carbon bas
ed NMs","Model developed by Xia et al in 2011","linearmodel")
```

```
Model with id: gSvjUZ17EEAV5OWL7Uls created. Please visit https://app.jaqpot.org/
```

PMML

```
from sklearn2pmml.pipeline import PMMLPipeline
from sklearn2pmml.pipeline import PMMLPipeline
pipelinepmmllinear = PMMLPipeline([
    ("scaler", MinMaxScaler()), ("MLR", LinearRegression())
])
pipelinepmmllinear.fit(X_train, Y_train)

from sklearn2pmml import sklearn2pmml
sklearn2pmml(pipelinepmmllinear, "log_k.pmml", with_repr = True)
```

```
                <Apply function="+">
                  <Apply function="*">
                    <FieldRef field="b"/>
                    <Constant dataType="double">1.923076923076923</Constant>
                  </Apply>
                  <Constant dataType="double">-0.1346153846153846</Constant>
                </Apply>
              </DerivedField>
              <DerivedField name="mix_max_scaler(V)" optype="continuous" dataType="double">
                <Apply function="+">
                  <Apply function="*">
                    <FieldRef field="V"/>
                    <Constant dataType="double">2.21729490022173</Constant>
                  </Apply>
                  <Constant dataType="double">-1.7184035476718407</Constant>
                </Apply>
              </DerivedField>
            </TransformationDictionary>
            <RegressionModel functionName="regression">
              <MiningSchema>
                <MiningField name="Log k" usageType="target"/>
                <MiningField name="R"/>
                <MiningField name="p"/>
                <MiningField name="a"/>
                <MiningField name="b"/>
                <MiningField name="V"/>
              </MiningSchema>
              <RegressionTable intercept="2.5057734675275714">
                <NumericPredictor name="mix_max_scaler(R)" coefficient="-0.12286883757006958"/>
                <NumericPredictor name="mix_max_scaler(p)" coefficient="1.3949141157229725"/>
                <NumericPredictor name="mix_max_scaler(a)" coefficient="-0.33960001214906327"/>
                <NumericPredictor name="mix_max_scaler(b)" coefficient="-2.0023541113199403"/>
                <NumericPredictor name="mix_max_scaler(V)" coefficient="3.112931796893343"/>
              </RegressionTable>
            </RegressionModel>
          </PMML>
```

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<PMML xmlns="http://www.dmg.org/PMML-4_3" xmlns:data="http://jpmml.org/jpmml-model/InlineTable" version="4.3">
  <Header>
    <Application name="JPMML-SkLearn" version="1.5.14"/>
    <Timestamp>2019-05-26T09:51:14Z</Timestamp>
  </Header>
  <MiningBuildTask>
    <Extension>PMMLPipeline(steps=[('scaler', MinMaxScaler(copy=True, feature_range=(0, 1))),
      ('MLR', LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False))])</Extension>
  </MiningBuildTask>
  <DataDictionary>
    <DataField name="Log k" optype="continuous" dataType="double"/>
    <DataField name="R" optype="continuous" dataType="double"/>
    <DataField name="p" optype="continuous" dataType="double"/>
    <DataField name="a" optype="continuous" dataType="double"/>
    <DataField name="b" optype="continuous" dataType="double"/>
    <DataField name="V" optype="continuous" dataType="double"/>
  </DataDictionary>
  <TransformationDictionary>
    <DerivedField name="mix_max_scaler(R)" optype="continuous" dataType="double">
      <Apply function="+">
        <Apply function="*">
          <FieldRef field="R"/>
          <Constant dataType="double">1.3679890560875512</Constant>
        </Apply>
        <Constant dataType="double">-0.8385772913816688</Constant>
      </Apply>
    </DerivedField>
    <DerivedField name="mix_max_scaler(p)" optype="continuous" dataType="double">
      <Apply function="+">
        <Apply function="*">
          <FieldRef field="p"/>
          <Constant dataType="double">1.5873015873015877</Constant>
        </Apply>
        <Constant dataType="double">-0.8253968253968256</Constant>
      </Apply>
    </DerivedField>
    <DerivedField name="mix_max_scaler(a)" optype="continuous" dataType="double">
      <Apply function="*">
        <FieldRef field="a"/>
        <Constant dataType="double">1.4285714285714286</Constant>
      </Apply>
    </DerivedField>
    <DerivedField name="mix_max_scaler(b)" optype="continuous" dataType="double">
```

Predictive model of TiO2 NPs damage on membrane cell by SMILES-based optimal descriptor and Monte Carlo technique (CORAL software)

https://app.jaqpot.org/model/nTJgb4Ss3zHIYZEcbg78

```
import pandas as pd
from jaqpotpy import Jaqpot


from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV


df = pd.read_csv('16_dataset.csv')


print(list(df))   # Prints the headers of all column


stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]


pipelinelinear = Pipeline(stepslinear) # define the pipeline object.


cross_val_score(estimator=pipelinelinear, X=X_train, y=Y_train, cv=5, n_jobs=-1)  #Per
froms a 5-fold cross validation


pipelinelinear.fit(X_train, Y_train)
print('Training score: ', pipelinelinear.score(X_train, Y_train))
print('Testing score: ', pipelinelinear.score(X_test, Y_test))
print('Total score: ', pipelinelinear.score(Xall, Yall))             #Trains the model
and prints R^2 statistics


jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")
jaqpot.set_api_key("……")


jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting LDH of TiO2 NPs","Mo
del developed by Toropova et al in 2013","linearmodel")
```

```
Model with id: nTJgb4Ss3zHIYZEcbg78 created. Please visit https://app.jaqpot.org/
```

Cytotoxicity of metal oxide to bacteria E.Coli models by Periodic table-based descriptors and stepwise-MLR

https://app.jaqpot.org/model/nTJgb4Ss3zHIYZEcbg78

```
import pandas as pd
from jaqpotpy import Jaqpot


from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV


df = pd.read_csv('29a_29b_dataset_full.csv')
df_train = pd.read_csv('29a_29b_dataset_mod1_training.csv')
df_test = pd.read_csv('29a_29b_dataset_mod1_test.csv')


print(list(df))   # Prints the headers of all column
print(list(df_train))
print(list(df_test))


Xall=df[['x', 'xox']]
X_train=df_train[['x', 'xox']] # Define the columns that will be used as independent f
eatures
X_test=df_test[['x', 'xox']]


Yall=df['pEC50']
Y_train=df_train['pEC50']    # Define the end-point
Y_test=df_test['pEC50']
stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]


pipelinelinear = Pipeline(stepslinear) # define the pipeline object.


cross_val_score(estimator=pipelinelinear, X=X_train, y=Y_train, cv=5, n_jobs=-1)  #Per
froms a 5-fold cross validation


pipelinelinear.fit(X_train, Y_train)
print('Training score: ', pipelinelinear.score(X_train, Y_train))
print('Testing score: ', pipelinelinear.score(X_test, Y_test))
```

```
print('Total score: ', pipelinelinear.score(Xall, Yall))        #Trains the model and
prints R^2 statistics

jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")

jaqpot.set_api_key(".....................")

jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting pEC50 in metal oxide
s, MLR, mod_1","Model developed by Kar et al in 2014, Strepwise MLR, mod_1","linearmod
el")
```

```
Model with id: QgRRwyU8r7e0NubEuDdX created. Please visit https://app.jaqpot.org/
```

Photo-induced toxicity of metal oxide NPs to E. Coli by MLR (dark condition case)
https://app.jaqpot.org/model/hygpzrH71XS1Wr8lGS69

```
import pandas as pd

from jaqpotpy import Jaqpot

from sklearn.linear_model import LinearRegression

from sklearn.preprocessing import MinMaxScaler

from sklearn.pipeline import Pipeline

from sklearn.model_selection import train_test_split

from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV

df = pd.read_csv('31_dataset_training_dark.csv')

print(list(df))   # Prints the headers of all column

Xall=df[['QMELECT', 'LZELEHHO']] # Define the columns that will be used as independent
features

Yall=df['-LOG(LC50)']    # Define the end-point

stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]

pipelinelinear = Pipeline(stepslinear) # define the pipeline object.

cross_val_score(estimator=pipelinelinear, X=Xall, y=Yall, cv=5, n_jobs=-1)  #Perfroms
a 5-fold cross validation


pipelinelinear.fit(Xall, Yall)

print('Training score: ', pipelinelinear.score(Xall, Yall))

#Trains the model and prints R^2 statistics


jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")


jaqpot.set_api_key("……")


jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting cytotoxicity for met
al oxides_dark","Model developed by Pathakoti et al in 2014, dark condition case","lin
earmodel")
```

```
Model with id: hygpzrH71XS1Wr8lGS69 created. Please visit https://app.jaqpot.org/
```

Photo-induced toxicity of metal oxide NPs to E. Coli by MLR hoto-induced (light) case)
https://app.jaqpot.org/model/5gCY316DzDh1Fdw4aigo

```
import pandas as pd
from jaqpotpy import Jaqpot
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV


df = pd.read_csv('31_dataset_training_light.csv')


print(list(df))   # Prints the headers of all column


Xall=df[['ALZLUMO', 'Cp']] # Define the columns that will be used as independent featu
res


Yall=df['-LOG(LC50)']    # Define the end-point


stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]


pipelinelinear = Pipeline(stepslinear) # define the pipeline object.


cross_val_score(estimator=pipelinelinear, X=Xall, y=Yall, cv=5, n_jobs=-1)  #Perfroms
a 5-fold cross validation


pipelinelinear.fit(Xall, Yall)
print('Training score: ', pipelinelinear.score(Xall, Yall))


jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")
jaqpot.set_api_key("…….")


jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting cytotoxicity for met
al oxides_photo induced_light","Model developed by Pathakoti et al in 2014, photo indu
ced_light case","linearmodel")
```

```
Model with id: 5gCY316DzDh1Fdw4aigo created. Please visit https://app.jaqpot.org/
```

Predicting metal oxide Nps toxicity to E. Coli cell line by MLR

https://app.jaqpot.org/model/OAiBYuee5PLJ7F580f2J

```
import pandas as pd
from jaqpotpy import Jaqpot
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, GridSearchCV, RandomizedSearchCV


df = pd.read_csv('56_dataset_full.csv')


df_train = pd.read_csv('56_dataset_training.csv')


print(list(df))   # Prints the headers of all column
print(list(df_train))
print(df)
Xall=df[['Z/r', 'DHme+']]
X_train=df_train[['Z/r', 'DHme+']] # Define the columns that will be used as independe
nt features
Yall=df['Obs. log 1/EC50']
Y_train=df_train['Obs. log 1/EC50']    # Define the end-point
stepslinear = [('scaler', MinMaxScaler()), ('MLR', LinearRegression())]
pipelinelinear = Pipeline(stepslinear) # define the pipeline object.


cross_val_score(estimator=pipelinelinear, X=X_train, y=Y_train, cv=5, n_jobs=-1)  #Per
froms a 5-fold cross validation


pipelinelinear.fit(X_train, Y_train)
print('Training score: ', pipelinelinear.score(X_train, Y_train))
print('Testing score: ', pipelinelinear.score(X_test, Y_test))
print('Total score: ', pipelinelinear.score(Xall, Yall))            #Trains the model
and prints R^2 statistics
jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")
jaqpot.set_api_key("……")
jaqpot.deploy_pipeline(pipelinelinear,Xall,Yall,"Model predicting metal oxide Nps toxi
city to E. Coli cell line by MLR","Model developed by Mu et al in 2016","linearmodel")
```

```
Model with id: OAiBYuee5PLJ7F580f2J created. Please visit https://app.jaqpot.org/
```

ΑΝΑΦΟΡΕΣ

1. 3Rs: Replacement, Refinement and Reduction of animals in research – DIRECTIVE 2010/63/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 September 2010 on the protection of animals used for scientific purposes

2. Andrew Ng. «Linear Regression - LMS algorithm». CS229 Lecture notes. Standford University. Pg. 4–7.

3. Breiman, Leo (1996). "Bagging predictors". Machine Learning. 24 (2): 123–140. doi:10.1007/BF00058655.

4. Buzea, Pacheco, & Robbie, Nanomaterials and nanoparticles: Sources and toxicity, 2007, Department of Physics, Queen's University, Kingston, Ontario K7L 3N6, Canada doi: 10.1116/1.2815690

5. Chomenidis et al, Jaqpot Quattro: A novel computational web platform for modelling and analysis in nanoinformatics, 2017, doi: 10.1021/acs.jcim.7b00223

6. Cros A.F.A. (1863) Action de l'alcool amylique sur l'organisme, Thesis, University of Strasbourg, Strasbourg, France.

7. Dubitzky, Werner; Granzow, Martin; Berrar, Daniel (2007). Fundamentals of data mining in genomics and proteomics. Springer Science & Business Media. p. 178.

8. Efron, B. (1979) "Bootstrap methods: Another look at the jackknife", The Annals of Statistics 7 (1): 1-26

9. Evaluation of the availability and applicability of computational approaches in the safety assessment of nanomaterials, Final report of the Nanocomput project, 2017

10. Ghosh Pallab, Introduction to Nanomaterials & Nanotechnology, Lecture 1, Department of Chemical Engineering IIT Guwahati, Guwahati–781039,India

11. Hervε Abdi, Partial Least Squares (PLS) Regression, The University of Texas at Dallas

12. Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pg. 278–282.

13. Hugo Kubinyi, Free Wilson Analysis. Theory, Applications and its Relationship to Hansch Analysis, 1988

14. John Dearden, The History and Development of Quantitative Structure-Activity Relationships (QSARs), 2016

15. Le Roux, Nicolas; Bengio, Yoshua; Fitzgibbon, Andrew (2012). «Improving First and Second-Order Methods by Modeling Uncertainty». Optimization for Machine Learning. MIT Press, pg. 404.

16. Marcu LG, Harriss-Phillips WM. In silico modelling of treatment-induced tumour cell kill: developments and advances. Comput Math Methods Med. 2012, doi: 10.1155/2012/960256

17. Meyer H. Zur Theorie der Alkoholnarkose Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung? Arch. Exp. Pathol. Pharmakol. 42, 109-18, 1899

18. MI Jordan, "Statistics and machine learning", 2014

19. Miramontes P. Un modelo de autómata celular para la evolución de los ácidos nucleicos [A cellular automaton model for the evolution of nucleic acids]. Tesis de doctorado en matemáticas. UNAM. 1992.

20. MoS2/TiO2 Heterostructures as Nonmetal Plasmonic Photocatalysts for Highly Efficient Hydrogen Evolution. Energy & Environmental Science

21. Nano, The Magazine for Small Science, What is Nanotechnology? A guide

22. Novel natural nanomaterial spins off from spider-mite genome sequencing. Phys.Org (May 23, 2013)

23. Overton E. (1901) Studien über die Narkose, Fischer, Jena, Germany,

24. Phil Simon (March 18, 2013). Too Big to Ignore: The Business Case for Big Data. Wiley, σελ. 89. ISBN 978-1-118-63817-0.

25. QSAR & QSPR, Alexandre Varnek, Faculté de Chimie, ULP, Strasbourg, FRANCE

26. Richet C. On the relationship between the toxicity and the physical properties of substances. Comptes Rendus Societe de Biologie 9, 775, 1893

27. S. Haykin. Νευρωνικά Δίκτυα και Μηχανική Μάθηση (3η Έκδοση)(Ε. Γκαγκάτσιου Μετάφραση), Παπασωτηρίου, 2010

28. Sarfaraz K. Niazi, Handbook of Preformulation, Chemical, Biological and Botanical Drugs, Second Edition, 2019, pg 105

29. Synthetic Nanomaterials Risk Assessment and Risk Management Basic report for the Swiss Action Plan, Environmental studies, Summary of the publication «Synthetische Nanomaterialien», Federal Office for the Environment FOEN and by the Federal Office of Public Health FOPH Bern, 2007, www.bafu.admin.ch/uw-0721-d

30. Todeschini, R., & Consonni Viviana. (2009). Molecular Descriptors for Chemoinformatics. (R.Mannhold, H. Kubinyi, & G. Folkers, Eds.) (2nd ed.). Wiley.

31. Wold S, Eriksson L (1995). "Statistical validation of QSAR results". In Waterbeemd, Han van de (ed.). Chemometric methods in molecular design. Weinheim: VCH. pp. 309–318.

32. Wold S., Sjostrom M. and Eriksson L., (2001). Pls-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, vol 58, pp. 109- 130.

33. Δημόπουλος, Β., Τσαντίλη-Κακουλίδου, Ά. 2015. Βασικές αρχές σχεδιασμού και ανάπτυξης φαρμάκων. Κεφ. 3. Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: https://repository.kallipos.gr/bitstream/11419/5884

34. Ελένη Βροντάκη et al, Ποσοτικές σχέσεις Δομής – Δράσης Τριών Διαστάσεων (3d - QSAr): Σύντομη Ανασκόπηση

35. Ι.Βλαχάβας, Π.Κεφαλάς, Ν.Βασιλειάδης, Φ.Κόκκορας, Η.Σακελλαρίου, Τεχνητή Νοημοσύνη, Εκδόσεις Γαρταγάνη, 2005

36. Κ. Γεωργούλη, ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ, Μια εισαγωγική Προσέγγιση Ελληνικά Ακαδημαϊκά Συγγράματα, 2015 http://repfiles.kallipos.gr/html_books/93/index.html

37. Καλαθάκης Χρήστος, Διπλωματική Εργασία ΔΙΑΓΝΩΣΤΙΚΗ ΑΕΡΙΟΣΤΡΟΒΙΛΩΝ ΜΕ ΧΡΗΣΗ ΜΗΧΑΝΩΝ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ (SUPPORT VECTOR MACHINES – SVM), Σχολή Μηχανολόγων Μηχανικών ΕΜΠ, 2010

38. Κυπαρισσίδης, Καμμώνα, Χαϊτίδου, 2008, Εφαρμογές Νανοτεχνολογίας στην Ιατρική Μια Υπόσχεση για το Μέλλον, Intellectum | Τεύχος 04 / Μάιος 2008

39. Λόκας Μάριος, Ταξινόμηση ανεπιθύμητης αλληλογραφίας εφαρμόζοντας στατιστικές τεχνικές ταξινόμησης με την γλώσσα προγραμματισμού R, Διπλωματική Εργασία, Τμήμα Πληροφορικής ΑΠΘ, 2012

40. Λούρου Σταυρούλα, Νανοτεχνολογία και εφαρμογές, Πτυχιακή Εργασία, Τμήμα Ηλεκτρονικής, ΤΕΙ Λαμίας, 2012

41. Μαγκανάρη Ειρήνη, ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΕ ΔΙΑΔΙΚΑΣΙΑ ΠΑΡΑΓΩΓΗΣ ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΒΙΟΜΗΧΑΝΙΑΣ ΠΑΡΑΓΩΓΗΣ ΤΣΙΜΕΝΤΟΥ, Διπλωματική Εργασία, Πανεπιστήμιο Πειραιά, 2006

42. Μαθηματικά και Στοιχεία Στατιστικής Γ' Ενιαίου Λυκείου, Οργανισμός Εκδοσεων Διδακτικών Βιβλίων

43. Μακρίδης Αντώνιος, In-vitro αξιολόγηση μαγνητικών νανοσωματιδίων ως φορέων μαγνητικής υπερθερμίας», Διπλωματική Εργασία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2013

44. Μπούτσικας Μιχαήλ. «Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)». Σημειώσεις μαθήματος "Στατιστικά Προγράμματα". Πανεπιστήμιο Πειραιώς

45. Παπαδόπουλος Ν. Αθανάσιος, Πρόβλεψη Τροχιών σε Δεδομένα Κίνησης με Βαθιά Νευρωνικά Δίκτυα, Διπλωματική Εργασία, Πανεπιστήμιο Πειραιώς, Δεκέμβριος 2018

46. Περρέα Δέσποινα, Εναλλακτικές Μέθοδοι, Ιατρική Σχολή Πανεπιστημίου Αθηνών.

47. Πετρίδης, Δ., 2015. Ανάλυση πολυμεταβλητών τεχνικών. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/2126

48. Τ. Σελλής, Τεχνητή Νοημοσύνη, Διάλεξη 10η, Νευρωνικά Δίκτυα - Μηχανική Μάθηση, 2007, Σχολή Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ Ε.Μ.Π.

49. Τσιάρα Αγγελική, Ταξινόμηση Εικόνων με Τυχαία Δάση, Μεταπτυχιακή Εργασία, Πανεπιστήμιο Ιωαννίνων, 2012

50. http://kelifos.physics.auth.gr/COURSES/neural/K8.pdf

51. http://www.business-analytics.gr/news/1211-machine-learning-vs-statistics

52. http://www.iep.edu.gr/images/IEP/EPISTIMONIKI_YPIRESIA/Epist_Monades/B_Kyklos/Genika/2017-10-31_Xhmeia_BLykeioy_nanoylika_ypodeigma1.pdf

53. https://egno.gr/2017/10/nanotechnologia-kafsimo-idrogono-apo-thalassino-nero-exagoi-ivridiko-nano-iliko-pou-kataskevase-erevnitis/

54. https://www.wikipedia.org/